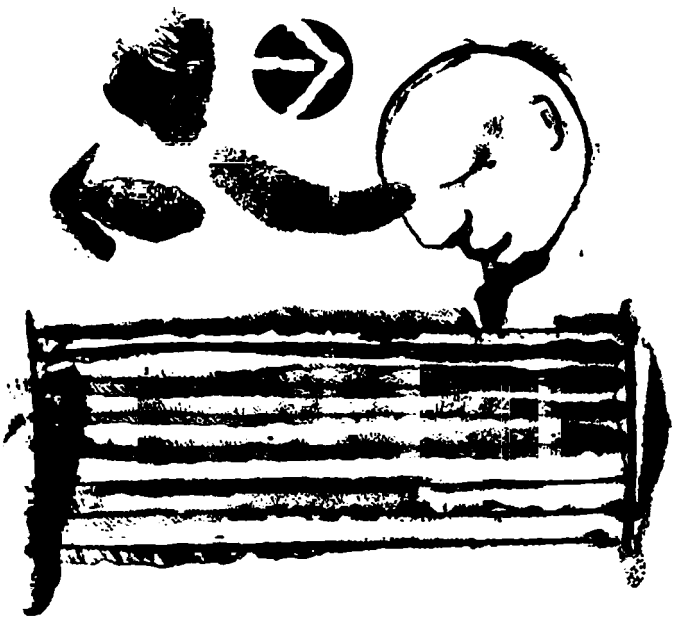
Ciencias Cognitivas

Pascual F. Martínez-Freire

La nueva filosofía de la mente



gedisa

La nueva filosofía de la mente

Este libro presenta lo que Pascual F. Martínez-Freire denomina "la nueva filosofía de la mente", es decir, la filosofía de la mente surgida tras la constitución de las ciencias cognitivas (básicamente psicología cognitiva e inteligencia artificial).

Esta obra ofrece las discusiones y temas básicos de la filosofía de la mente de una forma clara y autocontenida, de tal modo que para su comprensión no se requiere ningún conocimiento previo especializado. Además, el libro no se limita a la exposición de los autores más relevantes, sino que a lo largo de sus páginas se toma postura en cada tema, aportando comentarios personales, críticas y evaluaciones. Tras la discusión en los primeros capítulos de los autores clásicos, se dedica la mayor parte de la obra al estudio de los autores contemporáneos, por lo que el libro es en gran medida una historia de la filosofía actual. *La nueva filosofía de la mente* está destinada a un público amplio, incluyendo los estudiosos y estudiantes de filosofía, psicología o informática.

Pascual F. Martínez-Freire ha ejercido docencia regular en la Universidad Pontificia de Porto Alegre (Brasil), en la Complutense de Madrid y en la de Málaga. Ha sido profesor visitante en la Universidad de París IV, en la de Berkeley (California) y en la Sophia de Tokio (Japón). Actualmente es catedrático de Lógica y Filosofía de la Ciencia en la Universidad de Málaga y director del Título de Especialista en Ciencias Cognitivas Aplicadas. Es autor de libros de lógica y de filosofía de la ciencia y de numerosos artículos en diversas revistas científicas.

ISBN 84-7432-580-3

Código: 2.418

gedisa editorial

LA NUEVA FILOSOFÍA DE LA MENTE

por

Pascual F. Martínez Freire

En coedición con el Servicio de Publicaciones e Intercambio Científico de la Universidad de Málaga

gedisa
editorial

Diseño de la cubierta: Marc Valls

Primera edición, Septiembre de 1995, Barcelona

Derechos para todas las ediciones en castellano

© *by* Editorial Gedisa, S. A.
Muntaner, 460, entlo., 1.ª
Tel. 201 60 00
08006 - Barcelona, España

ISBN: 84-7432-575-7
Depósito legal: B. 32.195-1995

Esta obra fue realizada en coedición con el Servicio de Publicaciones e Intercambio Científico de la Universidad de Málaga

Impreso en Limpergraf
Calle del Río, 17, nave 3 - Ripollet

Impreso en España
Printed in Spain

Queda prohibida la reproducción total o parcial por cualquier medio de impresión, en forma idéntica, extractada o modificada, en castellano o cualquier otro idioma.

Indice

PREFACIO 13

1. Breve historia y temas de la filosofía de la mente 17
1. Primera aproximación a la noción de mente 17
2. Raíces antiguas en el estudio de la mente 18
3. El punto de partida clásico: Descartes 19
4. La posibilidad histórica de una filosofía de la mente basada en la ciencia: el comienzo de la psicología científica 20
5. El desarrollo de la filosofía de la mente 21
6. Temas principales de la filosofía de la mente 22

2. El problema *res cogitans-res extensa* en Descartes 25
1. El contexto metafísico del problema 25
2. El hombre como *res cogitans* 26
3. Interacción entre mente y cuerpo humanos 27
4. La unión del cuerpo con la mente 29

3. La crítica de Ryle al mito de Descartes 31
1. El mito cartesiano 31
2. El dualismo cartesiano como un error categorial 32
3. La naturaleza de los procesos mentales 34
4. El antimentalismo en la postura de Ryle 37
5. Conclusiones críticas 38

4. Mentalismo y antimentalismo en Wittgenstein ... 41
1. Wittgenstein y la filosofía de la mente ... 41
2. Filosofía de la mente y el argumento del lenguaje privado ... 43
3. Conductismo y mentalismo ... 46
4. La tarea de la psicología ... 49

5. Crítica del conductismo ... 53
1. Conductismo metodológico y conductismo ontológico ... 53
2. Breve presentación del conductismo ... 54
3. Valoración crítica del conductismo ... 58

6. Mentalismo y psicología popular ... 63
1. Mentalismo frente a conductismo ... 63
2. Caracteres de la psicología popular ... 64
3. Jerry Fodor y la psicología popular ... 66
4. Reservas y ataques a la psicología popular ... 69
5. Conclusiones ... 72

7. Teoría causal de la mente: materialismo y funcionalismo ... 77
1. La teoría causal de la mente ... 77
2. La teoría de la identidad mente-cerebro: Armstrong ... 78
3. La teoría de la identidad mente-cerebro: Lewis ... 82
4. Evaluación del materialismo ... 83
5. La invención del funcionalismo: Putnam ... 89
6. Microanálisis y análisis funcional en Fodor ... 94
7. Funcionalismo y el problema de los qualia ... 96

8. Inteligencia natural e inteligencia artificial ... 99
1. La cuestión de Turing ... 99
2. La inteligencia artificial y sus primeros críticos ... 102
3. Test de Turing y habitaciones orientales ... 112
4. Conclusiones ... 119

9. El problema mente-cerebro: datos e hipótesis ... 123
1. Introducción al problema ... 123
2. Las soluciones al problema mente-cerebro ... 125

3. El emergentismo de Searle 128
4. El dualismo de Eccles y su trasfondo metafísico 132
5. Fundamentos neurológicos del dualismo de Eccles 134
6. La hipótesis dualista interaccionista 136
7. Evaluación del dualismo de Eccles 139

10. Mentalismo y psicología cognitiva 143
1. Origen y desarrollo de la psicología cognitiva 143
2. Caracteres de la psicología cognitiva 145
3. Psicología cognitiva clásica 149
4. Psicología cognitiva conexionista 151

REFERENCIAS BIBLIOGRÁFICAS 155

INDICE DE AUTORES 163

Prefacio

La filosofía de la mente constituye una de las disciplinas filosóficas más pujantes en nuestros tiempos, interesando no sólo a las personas preocupadas por la filosofía, la psicología, la lingüística o la inteligencia artificial, sino también al público culto en general, en cuanto ofrece el estudio y discusión de una serie de temas que están en el centro de atención de las personas de nuestra época.

La filosofía de la mente, aunque tiene raíces antiguas, se ha consolidado como disciplina filosófica en torno a 1950, gracias a los trabajos de Gilbert Ryle y Ludwig Wittgenstein, entre otros. Sin embargo, lo que denomino "la nueva filosofía de la mente" surge a partir de 1977, cuando se consolidan de modo definitivo las llamadas "ciencias cognitivas". Las ciencias cognitivas componen un campo de investigación interdisciplinar cuyo tema aglutinador es el estudio del fenómeno de la cognición (conocimiento como recepción y manipulación de información), tanto en seres humanos como en animales y en máquinas. En este estudio están interesadas básicamente la psicología de orientación cognitiva y la ciencia de la inteligencia artificial, pero también la lingüística, la neurociencia, la informática y la lógica. La nueva filosofía de la mente se desarrolla teniendo en cuenta y valorando los descubrimientos de las ciencias cognitivas y, en particular, de la psicología cognitiva y de la inteligencia artificial.

El presente libro ofrece las discusiones y temas básicos de la filosofía de la mente de una forma clara y autocontenida, de tal modo que para su comprensión no se requiere ningún

conocimiento previo especializado. El libro está destinado al gran público, incluyendo los estudiantes y estudiosos de estos temas.

Por otra parte, el libro no se limita a la exposición de los autores relevantes en filosofía de la mente, sino que a lo largo de toda la obra se toma postura en cada tema aportando comentarios personales, críticas y evaluaciones.

Finalmente, la tercera característica de este libro consiste en que, tras la discusión en los primeros capítulos de los autores clásicos, se dedica la mayor parte de la obra al estudio de los autores actuales, razón por la cual el libro es en gran medida una historia de la filosofía actual.

En el primer capítulo se traza una breve historia de la filosofía de la mente, desde sus raíces antiguas hasta el surgimiento de las ciencias cognitivas, y asimismo se presentan los principales temas de la filosofía de la mente. El capítulo segundo está dedicado a Descartes, en cuanto constituye el punto de partida clásico en filosofía de la mente. Los capítulos tercero y cuarto estudian la constitución de la filosofía de la mente como disciplina filosófica, analizando la obra de Ryle y de Wittgenstein respectivamente.

Tras estos cuatro capítulos introductorios, el resto del libro se dedica a la nueva filosofía de la mente, es decir, al estudio de los problemas en torno a los procesos mentales en relación con las ciencias cognitivas de nuestra época. El capítulo quinto realiza una crítica del conductismo, crítica necesaria para comprender las razones de la psicología cognitiva actual. A su vez, el capítulo sexto completa al anterior al analizar la presencia del mentalismo en la llamada "psicología popular", mentalismo compartido por la psicología cognitiva, y al evaluar tal psicología popular.

El capítulo séptimo discute ampliamente la teoría causal de la mente, opuesta al conductismo, en sus dos variedades centrales: materialismo y funcionalismo. Se expone el materialismo o teoría de la identidad mente-cerebro en Armstrong y Lewis, haciendo una detallada evaluación del materialismo. Asimismo, se expone el funcionalismo en Putnam y Fodor, presentando las dificultades del funcionalismo.

En el capítulo octavo se caracteriza la inteligencia artificial, esto es, las capacidades de las máquinas inteligentes, comparándola con la inteligencia natural humana, discutiendo los argumentos a favor y en contra de la posibilidad de pensamiento en las máquinas, adoptando un punto de vista personal.

El capítulo noveno está dedicado al problema mente-cerebro, analizando las diferentes posturas generales. En particular se estudia y discute el emergentismo de John Searle, así como la hipótesis dualista interaccionista de John Eccles, que es objeto de una evaluación detallada.

Finalmente, el capítulo décimo se ocupa de caracterizar la psicología cognitiva, dentro del marco general de las ciencias cognitivas, señalando sus dos variantes básicas: la psicología computacional clásica y la psicología computacional conexionista.

Málaga, 12 de octubre de 1993

Pascual F. Martínez-Freire

1

Breve historia y temas de la filosofía de la mente

1. Primera aproximación a la noción de mente

Es de la mayor importancia resaltar que la mente, en oposición a una larga tradición filosófica, no es una substancia, es decir, no es una "cosa" individual, sino que es una colección de varios tipos de procesos, de tal manera que, inicialmente, podemos definir la mente como los procesos mentales.

Es evidente que tal caracterización tiene todo el aspecto de una definición circular, pero ello no debe importarnos en este comienzo ya que podemos realizar una descripción, aunque sea somera, de los principales procesos mentales, con lo que la explicación inicial de la mente puede resultar suficiente a estas alturas.

En efecto, entre los procesos mentales se encuentran las sensaciones cinestésicas (es decir, las relativas a nuestro propio cuerpo), las sensaciones acerca del mundo externo, las percepciones, los recuerdos, las imágenes mentales, los sentimientos, los deseos, las creencias, etcétera.

Utilizando conceptos tomados del punto de vista de la teoría de la información, podemos distinguir cuatro grupos principales de procesos mentales: 1) percepciones, esto es, recepción de información, 2) recuerdos o almacenamiento de información, 3) creencias, es decir, disposición y ordenación de la información, y 4) voliciones o disposición de la información con vistas a la ejecución de una acción. Por ejemplo, si vamos a una fiesta tendremos, entre otros, los siguientes

procesos mentales: percepciones acerca del lugar, sus muebles, alimentos y bebidas preparadas, las personas asistentes y sus nombres, etc., recordaremos algunos de los datos anteriores para su utilización posterior, formaremos creencias sobre movimientos en el lugar, así como sobre características de los alimentos y bebidas, cualidades de la gente, etc. y, también, estableceremos planes y objetivos acerca de acciones durante la fiesta o con motivo de la fiesta.

2. Raíces antiguas en el estudio de la mente

Como todos los temas de importancia, el asunto de la mente ya fue preocupación desde los tiempos antiguos. Por supuesto, entre los filósofos griegos podemos encontrar doctrinas de interés al respecto. La existencia de razón en los seres humanos, junto con ciertas ideas religiosas, condujo a los griegos a establecer una teoría del alma (psique).

Los dos principales filósofos de Grecia, Platón y Aristóteles, sostuvieron puntos de vista de influencia a lo largo de la historia al tiempo que enfrentados.

Platón (428-347) distinguió tres partes en el alma humana. Por un lado, el alma racional, caracterizada por su capacidad para pensar y para contemplar las Ideas (arquetipos eternos de las cosas). Por otro lado, el alma irascible, dotada de los sentimientos nobles, tales como la ambición o la valentía. Finalmente, el alma concupiscible, dotada con los sentimientos inferiores, como el placer o el instinto sexual. Para Platón se trata de tres partes que constituyen una unidad. Por ejemplo, en la obra platónica *Fedro* estas tres partes se comparan al conjunto formado por un auriga (alma racional) y los dos caballos (almas irascible y concupiscible) de su biga, los cuales ejecutan un esfuerzo único.

Pero en otros textos de Platón el alma auténtica es el alma racional (por ejemplo en el *Fedón*), que además es inmortal. Tal alma racional tiene que actuar sobre el mundo sensible, pero es enteramente diferente de él. Así pues, Platón ha defendido un dualismo de cuerpo y alma, que tuvo gran influencia en el mundo cristiano.

A su vez, Aristóteles (384-322), en su obra *De Anima*, concibe el alma (psique) como principio de la vida. En consecuencia hay tantos tipos de alma como tipos de vida. En primer lugar, tenemos el alma vegetativa, propia de las plantas, que es principio de crecimiento y de nutrición. En segundo lugar, tenemos el alma sensitiva, propia de los animales, dotada de las potencialidades de las plantas más las capacidades de movimiento y percepción. Y finalmente tenemos el alma intelectiva, propia de los hombres, que añade a las potencialidades de los animales la razón, tanto teorética como práctica.

Para Aristóteles todos los seres vivos tienen un alma, con lo que defiende cierto panpsiquismo. Por otra parte, el alma la entiende como forma del cuerpo, que es a su vez materia del alma, dentro de su concepción general hilemórfica. Por ello para este filósofo griego la psicología es una rama de la biología. Sin embargo, en *De generatione animalium*, señala que el alma inteligente (*nous*) puede ser "separable" del cuerpo. En todo caso, frente al dualismo de Platón, Aristóteles sostiene preferentemente un monismo de cuerpo y alma. Para él, las facultades psicológicas son poderes de los cuerpos y en los cuerpos.

Este breve examen de las concepciones platónica y aristotélica nos muestra que ya desde tiempos antiguos aparecen las dos posturas básicas acerca del problema mente-cuerpo, a saber, dualismo y monismo.

3. El punto de partida clásico: Descartes

Aunque desde la Grecia Antigua hasta el siglo XVII distintos autores han sostenido doctrinas acerca del alma y de su relación con el cuerpo, es Descartes el filósofo que planteó en toda su crudeza y dificultad el problema mente-cuerpo, de tal manera que sus consideraciones son calificadas como el punto de partida clásico en filosofía de la mente. En un capítulo posterior examinaremos en detalle las teorías cartesianas al respecto, limitándonos de momento a presentar su planteamiento general.

René Descartes (1596-1650) estableció la radical distinción entre pensamiento, atributo de la mente, y extensión, atributo del cuerpo, y al mismo tiempo sostuvo que el pensamiento carecía de extensión. Esta radical distinción implica dos espinosos problemas. Por un lado, el dualismo claro sobre mente y cuerpo conduce al problema de la explicación psicológica, es decir, a la posibilidad de una explicación de los procesos mentales que debe ser enteramente diferente de la explicación física. En segundo lugar, el dualismo mente-cuerpo plantea un conjunto de enigmas acerca de la relación entre mente y cuerpo, lo cual constituye propiamente el problema mente-cuerpo. Quiero decir que el problema mente-cuerpo no se plantea si no se sostiene algún tipo de dualismo mente-cuerpo o, lo que es lo mismo, que aquél no existe si tampoco se da el segundo.

Pues bien, ambos problemas, la explicación psicológica y la relación mente-cuerpo, están situados en el núcleo de la temática de la filosofía de la mente y, por ello, Descartes es considerado el punto de partida clásico.

4. La posibilidad histórica de una filosofía de la mente basada en la ciencia: el comienzo de la psicología científica

Posteriormente a Descartes también podemos encontrar doctrinas sobre la mente, como las de Spinoza o Leibniz, pero sólo el nacimiento de la psicología experimental hizo posible un tratamiento científico de los procesos mentales y, por tanto, una filosofía de la mente basada en consideraciones científicas.

Debe quedar claro que aquí, como en otros terrenos filosóficos, son realizables dos filosofías de corte bien distinto. Por un lado, la filosofía de la mente como una empresa enteramente especulativa, condicionada por ideas religiosas o por prejuicios materialistas o animistas. Y por otro lado, una filosofía de la mente como un conjunto de hipótesis basadas en datos científicos, no condicionada por religión o prejuicio alguno. Indudablemente es el segundo tipo de filosofía de la

mente el que merece respeto y estudio. Por ello es muy importante reseñar cómo y cuándo surge la psicología científica.

Los datos históricos relevantes son la creación en 1874 del laboratorio de Wilhelm Wundt y en 1876 del laboratorio de William James. Wundt (1832-1920) es reconocido como el fundador de la psicología experimental; su obra *Grundzüge der physiologischen Psychologie* (1873-1874) fue de hecho el primer libro de texto de psicología científica, y el laboratorio que estableció en Leipzig fue el primer laboratorio de psicología. A su vez, James (1842-1910) fue animado en sus estudios de psicología fisiológica por el conocimiento de la labor de Wundt mientras visitaba Alemania, estableciendo su propio laboratorio (peor dotado que el de Wundt) en Harvard. Así pues, a fines del siglo XIX la psicología se constituye como ciencia, saliendo del ámbito de la filosofía. La filosofía de la mente, desde entonces, ha de ser también una filosofía de la psicología.

5. El desarrollo de la filosofía de la mente

La filosofía de la mente, basada ya en los avances científicos de la psicología, se desarrolla con fuerza durante el siglo XX a través del Círculo de Viena y de la filosofía analítica.

El neopositivismo o Círculo de Viena, como es sabido, fue fundado por Moritz Schlick (1882-1936) en 1928, en la ciudad de Viena, cuando se constituye la Sociedad Ernst Mach y Schlick es nombrado su presidente. Este notable físico y filósofo agrupó en torno a él a varios pensadores de indudable mérito y preparación científica y filosófica. Entre ellos destaca Rudolf Carnap (1891-1970), quien se ocupó, entre otros muchos temas, de filosofía de la mente, en particular en su artículo *Psychologie in physikalischer Sprache* (1932-33).

Carnap defendió la reducción de la psicología a la física, dentro del fisicalismo general que propuso, es decir, la constitución de un lenguaje científico universal reducido al lenguaje propio de la ciencia física. Por otra parte, para los neopositivistas la filosofía de la mente no era sino un aspecto de la filosofía de la ciencia.

Dentro de la filosofía analítica cabe destacar, en primer lugar, la obra de Bertrand Russell (1872-1970) *The Analysis of Mind* (1921). En este libro Russell, además de examinar los principales conceptos psicológicos con gran agudeza, sostuvo su doctrina del monismo neutral; según ella, la mente y la materia no son dos tipos radicalmente diferentes de entidades, sino que ambas están construidas a partir de una misma estofa o material. El carácter metafísico de tal conjetura ha hecho del monismo neutral una pieza de museo sin vigencia actual.

Mucho mayor interés tiene la obra de Gilbert Ryle (1900-1976) titulada *The Concept of Mind* (1949), que estudiaremos en detalle más adelante. De momento diremos que en este trabajo Ryle se enfrenta decididamente al dualismo cartesiano, denunciándolo como un error categorial del lenguaje. Esta obra de Ryle constituye una obra clásica en la filosofía de la mente y su discusión es inevitable.

También son de gran importancia los estudios de Ludwig Wittgenstein (1889-1951) acerca de filosofía de la mente. Este tema, junto con la filosofía de la matemática, ocuparon los esfuerzos intelectuales del filósofo vienés durante la segunda época de su actividad filosófica (segundo Wittgenstein). Su obra más destacada al respecto es *Philosophische Untersuchungen* (1953). Aunque su publicación fue póstuma, Wittgenstein había revisado toda la obra y la había dejado prácticamente lista para ser publicada, cosa que no ocurre con los demás trabajos póstumos de este autor. Además, las ideas de Wittgenstein sobre filosofía de la mente ya eran conocidas por sus discípulos antes de la publicación de la citada obra.

En todo caso, y esto es lo que interesa a nuestra breve historia, lo cierto es que en torno a 1950 es cada vez más usual hablar de la filosofía de la mente como una actividad y disciplina filosófica diferenciada, gracias en particular a las contribuciones de Ryle y de Wittgenstein.

6. Temas principales de la filosofía de la mente

Al señalar los temas de la filosofía de la mente, debemos distinguir, y ésta es una de las tesis básicas de este libro, entre

lo que ha sido a lo largo de su historia la filosofía de la mente (desde sus momentos incipientes hasta 1977) y lo que es actualmente la filosofía de la mente (desde 1977) dentro del marco de las ciencias cognitivas. Es decir, debemos distinguir entre la filosofía de la mente anterior al desarrollo de las ciencias cognitivas y lo que llamaré la nueva filosofía de la mente.

Tradicionalmente la filosofía de la mente ha sido, en sus mejores momentos, filosofía de la psicología. Desde este punto de vista, son tres los temas generales de la filosofía de la mente, enumerados sin orden de prelación: 1) análisis de la explicación psicológica, 2) estudio de la naturaleza de los procesos mentales, y 3) el problema mente-cuerpo.

En efecto, y en primer lugar, la filosofía de la mente en cuanto aspecto de la filosofía de la ciencia se ocupa de analizar las peculiaridades de la explicación en psicología distinguiéndola, si hay lugar a ello, de la explicación en otras ciencias, en particular en la ciencia física.

En segundo lugar, la filosofía de la mente considera la naturaleza y características de los procesos mentales, ya tomados globalmente o bien analizando sus diferentes clases. Esta tarea se ha cumplido con dos estilos o formas distintas. Por un lado, análisis de los conceptos psicológicos tal como se presentan y funcionan en el lenguaje ordinario o también tal como los utiliza la llamada psicología popular (es decir, nuestro conocimiento ordinario de la conducta y motivaciones de las personas). Por otro lado, estudio de los fenómenos psicológicos a la luz de los descubrimientos de la psicología científica.

Finalmente, la filosofía de la mente aborda el problema clásico de la posible distinción entre mente y cuerpo y, en caso positivo, de la relación entre procesos mentales y procesos corporales. Las soluciones fundamentales actuales, como tendremos ocasión de analizar en detalle, son la teoría de la identidad entre procesos mentales y procesos cerebrales (materialismo), el dualismo o distinción radical entre lo mental y lo físico, el emergentismo (que sostiene que aunque lo mental no es independiente de los procesos cerebrales tampoco se reduce a éstos) y el funcionalismo, que define los procesos

mentales como ciertas funciones realizables (también) por los cerebros.

Sin embargo, cabe hablar de la nueva filosofía de la mente, encuadrada en el ámbito de las ciencias cognitivas. Dicho de modo rápido, las ciencias cognitivas constituyen un campo de investigación interdisciplinar cuyo tema aglutinador es el estudio del fenómeno de la cognición, tanto en seres humanos como en animales y en máquinas. En este estudio están interesadas básicamente la psicología de orientación cognitiva y la inteligencia artificial, pero también la lingüística, la neurociencia, la informática y la lógica. Aunque hay antecedentes en la constitución de este nuevo campo de investigación, podemos considerar que las ciencias cognitivas se consolidan definitivamente cuando en 1977 se crea la revista *Cognitive Science,* seguida dos años más tarde de la primera reunión de la Cognitive Science Society en La Jolla (California).Tras la consolidación de las ciencias cognitivas, la filosofía de la mente recibe nuevo impulso y nueva orientación.

La nueva filosofía de la mente, según entiendo, debe trabajar en tres niveles distintos: 1) como filosofía de las ciencias cognitivas, 2) como nueva teoría del conocimiento, y 3) como semántica filosófica. En cuanto al primer nivel, la filosofía de la mente ya no es solamente filosofía de la psicología sino también de la inteligencia artificial, siendo aquí su tema central la elucidación de la naturaleza de la mente en general y la discusión de la mente humana, la mente animal y las mentes mecánicas. A su vez, en cuanto nueva teoría del conocimiento, contrapuesta a la vieja teoría metafísica del conocimiento, su tema central es el fenómeno de la cognición en sus diferentes aspectos y variantes. Finalmente, como semántica filosófica debería denominarse teoría de la representación, ocupándose del análisis de los sentidos, significados y referencias. Si en el primer nivel la filosofía de la mente es parte de la filosofía de la ciencia y en el segundo nivel es la nueva filosofía del conocimiento, en este tercer nivel la filosofía de la mente es parte de la filosofía del lenguaje.

2

El problema *res cogitans-res extensa* en Descartes

1. El contexto metafísico del problema

René Descartes (1596-1650), en sus *Principia Philosophiae* (1644), establece una definición de substancia en términos muy exigentes, ya que dice que cuando concebimos la substancia concebimos únicamente una cosa (*res*) que existe de tal manera que no tiene necesidad más que de sí misma para existir. Así pues, para Descartes la substancia se caracteriza por la autosuficiencia, por ser una entidad radical. Por ello no resultará extraño que Spinoza, siguiendo a Descartes, sostenga que sólo Dios es substancia puesto que sólo la realidad divina posee auténtica autosuficiencia.

Ahora bien, en la filosofía cartesiana la substancia tiene atributos y tiene modos, siendo los primeros propiedades esenciales y siendo los segundos propiedades accidentales (color, forma, etc.). Por otro lado, aparte de la substancia infinita o Dios, existen según Descartes dos tipos de substancias, ambas finitas, que son la *res cogitans* y la *res extensa*. La primera se caracteriza por su atributo o propiedad esencial que es el pensamiento, careciendo en cambio de extensión. La *res cogitans* no ocupa lugar alguno y es simple, sin partes, y no es sino la mente. A su vez, la *res extensa* tiene como atributo o propiedad esencial la extensión, ocupa un lugar y es compleja, con partes, y no es sino el cuerpo.

Dada la radical oposición establecida por Descartes entre los atributos de pensamiento y de extensión, existe también

una radical oposición entre las mentes y los cuerpos, caracterizados respectivamente por tales atributos. Surge así el problema *res cogitans-res extensa*, es decir, la dificultad de establecer relaciones entre mentes y cuerpos en virtud de su disparidad como substancias. Tal problema, según Descartes, sólo surge en los seres humanos, ya que los animales y las máquinas no tienen mente. Para este filósofo la evidencia de que algo tiene mente es su posesión de lenguaje, pero animales y máquinas carecen de lenguaje. Un discípulo de Descartes, Gerauld de Cordemoy (1626-1684), en su obra *Discours physique de la parole* (1677), va aún más lejos al sostener que el lenguaje es la propiedad específica de los seres humanos.

Así pues, el problema *res cogitans-res extensa* consiste en la dificultad de explicar la posibilidad y la realidad de las relaciones entre la mente y el cuerpo humanos, dada la radical oposición entre ambos tipos de realidades. El mismo Descartes señala que la propia experiencia plantea tal problema, ya que tenemos sensaciones, sentimientos y apetitos (de hambre, sed, etc.) que no pertenecen ni sólo a la mente ni sólo al cuerpo, sino a la unión entre cuerpo y mente. En consecuencia, y ateniéndonos a las distinciones metafísicas cartesianas, tenemos un dualismo radical entre mente y cuerpo, junto con la necesidad de explicar las relaciones entre cuerpo y mente en el ser humano.

2. El hombre como *res cogitans*

Las dificultades se hacen mayores al encontrar textos cartesianos en los que se asevera que nosotros nos conocemos como mentes, como *res cogitans*. En efecto, en el *Discours de la méthode* (1637), Descartes dice que conoce que él es una substancia cuya esencia o naturaleza entera no es sino pensar y que para ser no tiene necesidad de ningún lugar ni depende de ninguna cosa material. Más aún, sigue Descartes, el yo, es decir el alma, por la cual es lo que es, es enteramente distinta del cuerpo, es más fácil de conocer que el cuerpo, y aunque éste no existiese el alma no dejaría de ser todo lo que es. Este texto, en la cuarta parte del *Discours*, viene a continuación de la

tesis cartesiana de que la proposición "pienso, luego existo" es el primer principio de la filosofía. Por ello no es de extrañar la reducción del yo al pensamiento.

En consecuencia, hay una distinción real entre la mente y el cuerpo del hombre, con lo que la relación entre mente y cuerpo humanos se torna aún más problemática. Pero al mismo tiempo, y puesto que el hombre se conoce como *res cogitans*, el dualismo parece desvanecerse; cada ser humano es únicamente su alma o mente y su cuerpo es ajeno a él mismo.

No tengo especial interés en salvar a Descartes de esta contradicción, a saber, que, por un lado, el hombre es una realidad dual de mente y cuerpo mientras que, por otro lado, el hombre es únicamente su mente, pero en todo caso parece claro que el filósofo francés no puede establecer de modo legítimo que el ser humano es solamente mente. En efecto, del hecho de que Descartes, o cualquier otro hombre, pueda imaginarse como reducido a su mente no se sigue que el ser humano sólo es *res cogitans*. Además nuestro autor, en otros textos, sostiene la interacción entre cuerpo humano y mente humana, con lo que se retoma el dualismo asumido por razones metafísicas, e incluso en otros textos se habla de cierta unión entre cuerpo y mente, con lo que se matiza el dualismo. En suma, el dualismo cartesiano no es una tesis tan clara como pudiera parecer a simple vista.

3. Interacción entre mente y cuerpo humanos

El filósofo francés, en *Les Passions de l' âme* (1649), distingue tres tipos de procesos humanos. En primer lugar, procesos corporales y mecánicos, como la digestión o los reflejos musculares, que sólo dependen del cuerpo. En segundo lugar, procesos en los que colaboran cuerpo y mente, tales como las sensaciones, sentimientos o el conocimiento de otros cuerpos. Y finalmente, procesos que sólo dependen de la mente, como la percepción de las ideas claras y distintas. Por tanto, la relación entre mente y cuerpo humanos se plantea en el segundo tipo de procesos del hombre, no en todos los

procesos, habiendo, podemos decir, algunos estrictamente corporales y otros que son propiamente mentales.

En la Antigüedad, Aristóteles pensaba que el corazón era la sede de la mente mientras que Hipócrates ya había defendido que el cerebro era el lugar de la mente; posteriormente Galeno (siglo II después de Cristo) siguió el punto de vista de Hipócrates y sostuvo además que los nervios eran conductos huecos por los que fluyen los *spiritus animales* (una especie de soplos de animación o activación) pasando mensajes entre el cerebro y los músculos.

Descartes mantuvo la teoría de los *spiritus animales* e insistió en la importancia del cerebro. De hecho la teoría de los *spiritus animales* fue sustentada hasta el siglo XVIII, cuando Luigi Galvani (1737-1798) descubrió la naturaleza eléctrica de la transmisión nerviosa.

Para Descartes, la mente, aunque es incorporal, interactúa con el cuerpo humano en la glándula pineal. Tal órgano (también llamado epífisis) se encuentra en el centro del cerebro medio y cabe pensar que nuestro filósofo le atribuyó el papel de punto de encuentro entre la mente y el cuerpo humanos a causa de su posición central en el cerebro y porque, erróneamente por cierto, Descartes creía que los animales no tenían glándula pineal. (Hoy día no se atribuye un papel preeminente a la glándula pineal, aunque se cree que tiene que ver de algún modo con la regulación sexual).

Así pues, según Descartes, es en la glándula pineal donde los *spiritus animales* son movidos por la mente y a su vez la mente es influida por tales *spiritus*. Tal como dice en *Les Passions de l'âme*, estos espíritus animales son las partes más vivas y más sutiles de la sangre, las cuales entran en las cavidades del cerebro, llegan hasta los nervios y mediante éstos hasta los músculos, de tal manera que mueven el cuerpo de todas las maneras que puede ser movido. Por tanto, Descartes presenta al cuerpo y a la mente humanos como dos entidades que, aunque distintas, se influyen entre sí, ofreciendo así un dualismo interaccionista.

4. La unión del cuerpo con la mente

A pesar de la distinción, por razones metafísicas, entre mente y cuerpo humanos, Descartes defendió asimismo la unión entre ellos, esto es, que la mente está unida a todo el cuerpo. Tal tesis evidentemente contraría las consideraciones anteriores, ya que de la interacción entre mente y cuerpo en la glándula pineal pasamos a la doctrina, de abolengo escolástico, de la unión entre cuerpo y mente.

Esta tesis aparece en las *Meditationes de Prima Philosophia* (1641), donde dice que la naturaleza le enseña que no está solamente alojado en su cuerpo, como un piloto en su navío, sino que está tan estrechamente y de tal manera confundido y mezclado con su cuerpo que compone como un solo todo con él; e incluso habla explícitamente de la unión y como mezcla del espíritu con el cuerpo. En esa misma obra, en respuesta a una objeción de Arnauld, Descartes declara que el espíritu y el cuerpo son substancias incompletas en relación con el hombre que componen, aunque espíritu y cuerpo considerados por separado son substancias completas. Desde este punto de vista podríamos decir, en una tentativa por aclarar la filosofía cartesiana, que mente y cuerpo serían distintos considerados abstractamente pero que están unidos considerados realmente en el ser humano. Sin embargo, entonces la substancia no sería ni el cuerpo ni la mente sino el propio ser humano. En todo caso, aquí la distinción *res cogitans-res extensa* es disminuida, ya que en el hombre ambas substancias se necesitan entre sí para componer el hombre. Con ello el dualismo metafísico entre *res cogitans* y *res extensa* es reducido al analizar la composición del ser humano.

Para mayor confusión, en *Les Passions de l'âme*, Descartes afirma sucesivamente, en los artículos 30 y 31, que el alma está unida a todas las partes del cuerpo conjuntamente y, también, que el alma ejerce sus funciones de modo particular en la glándula pineal.

En conclusión, el dualismo cartesiano no es una tesis clara, ya que Descartes sostiene respecto del asunto tres posturas distintas: radical distinción entre cuerpo y mente

humanas, al tiempo que interacción entre ambos (dualismo fuerte interaccionista), reducción del hombre a su mente (monismo espiritualista) y unión estrecha entre cuerpo y mente humanas (monismo animista o hilemorfismo).

De todas formas, los principales errores de Descartes en filosofía de la mente son los tres siguientes. En primer lugar, al concebir la mente como una substancia, dejó la cuestión en el terreno abstracto de la metafísica, no estimulando la necesaria investigación empírica. En segundo lugar, su idea de que una realidad extensa no puede pensar contraría la evidencia empírica de que el cerebro, que es extenso y real, efectivamente piensa. Y finalmente, Descartes rechazó la posibilidad de una mente mecánica, ya que el ámbito del mecanicismo se reduce a los cuerpos extensos y las mentes no son extensas, y sin embargo hoy en día tenemos la evidencia empírica de máquinas que piensan (como los ordenadores que juegan al ajedrez).

3

La crítica de Ryle al mito de Descartes

1. El mito cartesiano

En 1949 Gilbert Ryle(1900-1976), filósofo analítico del grupo de Oxford, publicó un libro, *The Concept of Mind*, que supuso un fuerte impulso a la filosofía de la mente.

En esta obra, Ryle sostiene que Descartes dejó como legado la teoría que defiende el dualismo entre mente y cuerpo. Tal teoría, a los ojos del filósofo inglés, constituye una especie de *doctrina oficial*, que está incorporada en el lenguaje y que deriva asimismo de los teólogos escolásticos y de la Reforma. Más aún, tal dualismo cartesiano es un *dualismo popular*. Según tal dualismo, el cuerpo es espacial y observable, sujeto a las leyes mecánicas, mientras que la mente no es espacial ni observable (sino privada), siendo independiente de las leyes de la mecánica.

En cada hombre hay, según el dualismo cartesiano, dos historias paralelas: la historia física y pública del cuerpo, y la historia privada y no-física de la mente. O dicho de otro modo, en cada hombre hay una existencia física y una existencia mental. En consecuencia, insiste Ryle, las conexiones entre los episodios de la vida privada y los episodios de la vida pública son misteriosas. Tal teoría cartesiana es denominada por Ryle "el dogma del fantasma en la máquina" (es decir, la mente inobservable en la máquina del cuerpo), y además constituye un "mito filosófico", el mito cartesiano.

Como primeras observaciones a los cargos de Ryle contra

Descartes podemos señalar lo siguiente. Para empezar, es muy notable que las aseveraciones del filósofo inglés acerca del filósofo francés no van acompañadas de cita alguna de las obras cartesianas, lo cual hace sospechar que se trata de inventar un autor más que de estudiarlo. En segundo lugar, y tal como hemos visto, cuando Descartes propone el dualismo (y no es su única postura) tal dualismo es interaccionista y no se trata de un paralelismo (tal como, en cambio, encontramos en Leibniz). En tercer lugar, Descartes rechaza explícitamente la tesis del "fantasma en la máquina", al decir literalmente que el alma no está alojada en el cuerpo como un piloto en su navío. Y finalmente, en última instancia, tal como veremos más adelante, el ataque de Ryle a Descartes es contra el mentalismo, es decir, contra la tesis de que los procesos mentales son internos y privados, no necesariamente manifestados en la conducta; con ello Ryle se mostrará muy próximo a la psicología conductista, defendiendo un conductismo filosófico.

2. El dualismo cartesiano como un error categorial

Para Gilbert Ryle la filosofía es el esfuerzo para sustituir hábitos categoriales (inapropiados) por una disciplinada categorización. Tal concepción de la filosofía se enmarca en el ámbito de la filosofía analítica como filosofía lingüística o, más exactamente, dentro de la consigna del segundo Wittgenstein de reordenar el uso del lenguaje.

En efecto, para nuestro autor hablar de la mente como algo diferente del cuerpo y al mismo tiempo semejante a él en cierto sentido es un error categorial. El dualismo cartesiano, según Ryle, considera que, como el cuerpo humano es una unidad compleja organizada, la mente humana también debe ser una unidad compleja organizada, aunque constituida por elementos y estructura diferentes. Es decir (aclarando el punto de vista ryleano) Descartes creería que la mente es distinta del cuerpo pero sometida a causas y efectos al igual que el cuerpo. Los ejemplos de Ryle intentan aclarar el error

categorial de, digamos, tomar a la mente como otro tipo de cuerpo aunque diferente del cuerpo.

Supongamos que un extranjero visita Oxford por primera vez y se le muestran los *colleges*, bibliotecas, campos de deporte, museos y oficinas administrativas, y, tras todo ello, pregunta dónde está la Universidad de Oxford. Entonces debemos explicarle que la Universidad no es un edificio más, comparable a los que ya ha visto, sino la manera en que todo lo que ha visto se encuentra organizado. Tal extranjero comete el error categorial de confundir la Universidad con algún edificio aún no visto del tipo de los edificios que ha visto. Asimismo comete error categorial el niño que observa el desfile de una división y que, tras señalársele los batallones, baterías y escuadrones, pregunta cuándo va a desfilar la división, suponiendo (erróneamente) que la división es similar a las unidades que ha visto pasar aunque diferente de ellas. Entonces se le mostraría su error diciendo que al ver pasar los batallones, baterías y escuadrones ya estaba viendo desfilar la división.

Análogamente, podemos añadir para aclarar la exposición de Ryle, es un error categorial creer que la mente es algo tan substancial como el cuerpo aunque distinto del cuerpo. Para el filósofo inglés, los procesos mentales no son realidades fantasmagóricas que existen separadamente del cuerpo, pero los procesos mentales tampoco son exactamente procesos corporales (al igual que la Universidad no existe al margen de sus edificios pero tampoco es un edificio más). Como veremos a continuación, para Ryle los procesos mentales son disposiciones a la conducta.

En esta línea de argumentación del error categorial nuestro autor extrae algunas consecuencias. En primer lugar, la tradicional oposición entre mente y materia queda diluida, aunque de un modo distinto a como se la diluye en las también tradicionales reducciones de la materia a la mente o de la mente a la materia. En segundo lugar, queda claro que tanto el idealismo (que reduce el mundo material a procesos y estados mentales) como el materialismo (que reduce los procesos mentales a estados y procesos físicos) son respuestas

inadecuadas. Nos convendrá recordar esta oposición de Ryle al materialismo.

3. La naturaleza de los procesos mentales

Para Gilbert Ryle la expresión "mi mente" no se refiere a órgano alguno, es decir, no existe un órgano especial (corporal o incorporal) que podamos denominar "mente". En todo caso no siempre es posible evitar el uso de la palabra "mente", pero cuando se nos exige pureza lógica debemos seguir, según nuestro autor, el ejemplo que nos ofrecen los novelistas, biógrafos y escritores de diarios personales, quienes hablan de personas que hacen o experimentan cosas. Así pues, de algún modo la palabra "persona" puede sustituir a la palabra "mente". Sin embargo, lo definitivo para Ryle es que mi mente quiere decir mi aptitud y disposición a hacer determinado tipo de cosas, y no significa una especie de aparato personal sin el cual no podría hacerlas.

Por tanto, podemos comentar, la mente no es una substancia sino que la mente son procesos mentales, y esto es correcto tal como señalé en el principio de este libro. Pero además, para nuestro autor, los procesos mentales no son propiedades actuales sino disposiciones para la conducta, y aquí encontraremos problemas.

Para Ryle, no todas las disposiciones son relevantes cuando pretendemos hablar acerca de la mente. Los términos disposicionales (significativos de disposiciones) que nos conciernen son, en primer lugar, aquellos apropiados para la caracterización de seres humanos. Más aún, la clase relevante de términos disposicionales es todavía más limitada, ya que se refiere a aquellos términos apropiados para caracterizar aspectos de la conducta humana que exhiben cualidades de intelecto y de carácter. No nos interesan, pues, reflejos que pudieran ser peculiares de los hombres ni elementos fisiológicos propios de la anatomía humana. Así pues, debemos señalar que, en Ryle, la mente es excluida de los animales (y *a fortiori* de las máquinas), al igual que ocurría en Descartes, siendo referida a capacidades humanas en relación con la

conducta inteligente y con la conducta moral, no con la conducta en general.

Expresando la mente disposiciones a la conducta inteligente o moral, debemos caracterizar antes los términos disposicionales en sentido amplio, tal como hace este filósofo. Ejemplos de términos disposicionales son "frágil" o bien "soluble". Cuando decimos que un objeto de vidrio es frágil no pretendemos que en un momento dado tal objeto deba romperse, ya que puede ser frágil sin que se haya roto; lo que queremos decir es que si alguna vez tal objeto es golpeado entonces se hará añicos. Asimismo, cuando decimos que el azúcar es soluble queremos indicar que si sumergimos un trozo de azúcar en un líquido entonces se disolverá. Por tanto, poseer una propiedad disposicional (disposición) no consiste, señala Ryle, en encontrarse en un estado particular o experimentar determinado cambio, sino ser susceptible de encontrarse en un estado particular o de experimentar un cambio cuando se cumple cierta condición. Así pues, a estas alturas podemos precisar que según el filósofo inglés los procesos mentales no son estados conductuales ni cambios corporales, sino disposiciones a la conducta.

Dejando los ejemplos generales de términos disposicionales, podemos seguir a Ryle en un ejemplo ya específicamente psicológico. Consideremos el talento de un payaso. Tal talento se muestra en su comportamiento visible, porque es el ejercicio de una habilidad. Ahora bien, una habilidad no es un acto, con lo que, dice literalmente nuestro autor, no es algo observable ni tampoco no observable; pero esto no se debe a que sea un acontecimiento oculto o fantasmal, sino a que no es un acontecimiento. El talento del payaso es una disposición o complejo de disposiciones, y una disposición, añade Ryle, es un factor de tipo lógico tal que no puede ser visto o no visto, grabado o no grabado.

Así pues, podemos comentar, los procesos mentales en cuanto disposiciones se manifiestan en la conducta, aunque no aparecen por sí mismos. Pero, desde este punto de vista, no podemos dejar de advertir el extraño carácter que a los procesos mentales atribuye Ryle, ya que son entidades ni observables ni no-observables.

En todo caso, nuestro filósofo intenta aclarar la naturaleza de los procesos mentales como disposiciones a la conducta distinguiendo entre saber-que y saber-cómo (*knowing that* y *knowing how*). Tal distinción señala la diferencia entre hechos y disposiciones. En efecto, mientras el saber-que se refiere al conocimiento teorético, el saber-cómo hace referencia al conocimiento de reglas y criterios; pero el conocimiento teorético es acerca de hechos, en cuanto que el conocimiento de reglas y criterios supone disponer de un conjunto de disposiciones. El saber-cómo, dice Ryle, no es una disposición simple, como un reflejo o un hábito, sino que su actualización comprende la observancia de reglas o cánones o la aplicación de criterios. Nuestro autor denuncia lo que califica de "leyenda intelectualista", esto es, la creencia en la superioridad del conocimiento teorético, o conocimiento acerca de hechos, sobre el saber-cómo. Pero además, añade, cuando se predican de una persona epítetos psicológicos tales como astuto, tonto, prudente o imprudente, no se le atribuye conocimiento o ignorancia de alguna verdad sino la habilidad o ineptitud para hacer cierto tipo de cosas.

Así pues, según Ryle, al ir más allá del comportamiento mismo no está tratando de aprehender un comportamiento oculto que es la contrapartida del primero y que acaece en un escenario secreto de la vida interna del sujeto, sino que considera las aptitudes e inclinaciones (en suma, disposiciones) de las cuales su conducta es una actualización. Hablar de la mente de una persona, dice este autor, no es hablar de un depósito en el que se pueden alojar objetos que nos está prohibido alojar en otro depósito llamado "mundo físico"; es más bien hablar de las aptitudes, debilidades y propensiones de esa persona para hacer y padecer determinados tipos de cosas en el mundo de todos los días.

En resumen, podemos decir, para Ryle los procesos mentales no son procesos conductuales aunque están manifestados en la conducta. Sin embargo, subsiste el problema de señalar la ubicación de las disposiciones, ya que nuestro autor no considera en modo alguno el papel del cerebro o del sistema nervioso en general. Y además, en cuanto a la contraposición

entre conocimiento teorético y disposiciones, Ryle no puede tomarla como ilustrativa de la naturaleza de los procesos mentales, so pena de negar que teorizar es un proceso mental.

4. El antimentalismo en la postura de Ryle

El objetivo último de la filosofía de la mente de Gilbert Ryle es eliminar el "mundo mental". Para él, no existen "episodios privados" ya que éstos pertenecerían al fantasma en la máquina, el cual resulta inadmisible. En consecuencia, tampoco existe un acceso privilegiado de cada uno a su propia mente: en ninguno de los sentidos en que comúnmente apreciamos si una persona sabe, o no, algo acerca de sí misma, es necesario o útil postular un "acceso privilegiado" para explicar cómo es que ha obtenido o podría haber obtenido tal conocimiento. Llego a valorar, dice también Ryle, la habilidad y las tácticas de un jugador de ajedrez observando su juego y el de otros jugadores, y el que yo mismo sea el jugador no marca ninguna diferencia.

En esta línea antimentalista, el filósofo inglés condena la introspección. Nos recuerda que la introspección se considera un supuesto tipo de observación, consistente en la observación por cada persona de sus estados o procesos mentales en cuanto escrutados deliberadamente y con atención, con la particularidad de que sólo el poseedor de tal estado o proceso puede observarlo y de que tal observación es infalible. Para Ryle, no existe tal percepción interna, distinta de la percepción sensorial, por la sencilla razón de que no existen los eventos internos que serían su objeto. Pero además, tal percepción interna, en caso de existir, requeriría que el observador pudiera prestar atención a dos cosas al mismo tiempo, lo cual no es posible según nuestro autor; por ejemplo, estaría a la vez atendiendo a un programa de preguntas y respuestas y también prestando atención a su acto de atender al programa. Finalmente, Ryle recuerda la objeción de David Hume (1711-1776) al carácter infalible de la introspección: algunos estados mentales no pueden ser escrutados fríamente debido al hecho de que encontrarnos en ellos implica estar envueltos en ellos, tales como los estados de pánico o de ira.

Estas críticas a la introspección, podemos comentar, pasan por alto el hecho mismo de la existencia de procesos de introspección, que Ryle se vé obligado a aceptar de algún modo al reducir la introspección a retrospección o recuerdo; pero también el recuerdo exige la existencia de eventos mentales internos. El malabarismo doctrinal de este autor le lleva a perder el equilibrio cuando dice que los objetos de nuestras retrospecciones, aunque personales, no es necesario que sean, pero pueden serlo, aspectos privados. Por otra parte, es posible atender a dos tareas al mismo tiempo y, además, la caracterización de la introspección como infalible no es una exigencia aceptable, con lo que no cabe argumentar en base a la ausencia de infalibilidad.

El antimentalismo de Ryle es coherente con su simpatía por el conductismo, la cual manifiesta al final de su libro, aunque declara que no ha tenido la finalidad de proponer a la psicología una metodología especial ni tampoco de criticar las hipótesis de tal o cual ciencia. En un capítulo posterior analizaremos el conductismo, pero podemos señalar ya que el conductismo, fundado por John Watson (1878-1958) con la publicación de su artículo "Psychology as the behaviourist views it" (1913), sostiene que los datos psicológicos deben referirse únicamente a la conducta públicamente observada de los organismos vivos (incluidos los animales), desechando la introspección y negando la existencia de una "vida mental" al margen de la conducta. Ryle, a su vez, declara que el plan metodológico del conductismo ha tenido una influencia revolucionaria para el plan de la psicología y, más aún, ha sido una de las fuentes de la sospecha filosófica de que la teoría de los dos mundos (físico y mental) es un mito.

5. Conclusiones críticas

Una primera conclusión acerca de la filosofía de la mente de Ryle es que nuestro autor no tiene en cuenta la psicología científica. En efecto, él mismo confiesa que a lo largo de su libro ha dicho muy poco acerca de la psicología, y que su obra puede describirse como un ensayo de psicología filosófica, no

de psicología científica. Los conceptos analizados, como aprendizaje, intento, fingir, querer o estar perturbado, no son, según Ryle, conceptos técnicos sino cotidianos. Su opinión acerca de la psicología como ciencia no es muy alta, ya que estima que debe abandonarse la idea de que la "psicología" es el nombre de una investigación única o de un conjunto de investigaciones íntimamente conectadas, para aceptar que denota una federación fortuita de investigaciones y técnicas. Más aún, para Ryle la verificación de las aptitudes e inclinaciones mentales de una persona es un proceso inductivo que no se lleva a cabo en el laboratorio o ayudados por métodos estadísticos. En suma, tenemos aquí un caso de filosofía de la mente elaborada sobre el análisis del lenguaje cotidiano, y no teniendo en cuenta la psicología científica.

Además de este error de procedimiento, Ryle incurre en otros errores de carácter doctrinal. En primer lugar, el rechazo de los procesos internos constituye una renuncia a la explicación de la conducta, ya que, como resulta evidente desde el punto de vista de la psicología cognitiva actual (de la que nos ocuparemos al final de este libro), la conducta requiere una explicación en términos de los procesos internos que la producen, ya sea una explicación específicamente psicológica o una explicación neurológica (si ésta llega a ser suficiente y relevante). De hecho, Ryle no se molesta en considerar el papel del cerebro en los procesos mentales, siendo en este punto inferior al propio Descartes a quien tanto critica.

En segundo lugar, tal como pusimos de relieve, la naturaleza de los procesos mentales como disposiciones a la conducta no queda debidamente aclarada, resultando extraño el carácter de las disposiciones como entidades que no son ni observables ni no-observables.

Y finalmente, frente a lo que parece la disyuntiva de Ryle, es decir, dualismo o antimentalismo, no tenemos que elegir exclusivamente entre ambos. Es perfectamente posible dejar de lado el antimentalismo sin adoptar el dualismo cartesiano, ya que si nos adherimos al mentalismo, adoptando un programa de investigación sobre las causas internas de la conducta, no estamos obligados a abrazar el espiritualismo, ya que los

procesos internos pueden ser considerados incluso en su totalidad procesos cerebrales (aunque sobre este punto nos extenderemos en un capítulo ulterior).

4

Mentalismo y antimentalismo en Wittgenstein

1. Wittgenstein y la filosofía de la mente

Ludwig Wittgenstein (1889-1951) es sin duda uno de los filósofos más originales e influyentes del siglo xx. Sus preocupaciones intelectuales fueron muy amplias, desde la metafísica y la ética a la lógica y la matemática pasando por el lenguaje y la psicología. Pero además su honestidad intelectual le llevó a una marcada evolución en su pensamiento, hasta el punto de que es legítimo (y habitual) hablar de un primer Wittgenstein y de un segundo Wittgenstein para indicar los dos períodos diferentes de su actividad filosófica.

Cada uno de estos períodos va unido a un libro distinto de nuestro filósofo. El primer Wittgenstein, en efecto, es el autor del *Tractatus logico-philosophicus* (1921) mientras que el segundo es el autor de *Philosophical Investigations* (1953), obra aparecida tras su muerte pero que él mismo había preparado y revisado para su publicación. También pueden darse fechas para situar ambos períodos. El primero iría desde 1912, cuando se incorpora como estudiante al Trinity College de Cambridge, hasta 1921, cuando tras la publicación del *Tractatus* abandona la actividad filosófica. A su vez, el segundo período iría desde 1929, fecha en que Wittgenstein (según su propio testimonio) volvió a escribir filosofía, hasta 1951, fecha en que falleció de cáncer.

En cuanto a una breve descripción del diferente contenido filosófico de ambas fases del pensamiento de Wittgenstein

podemos resumirla como sigue. Mientras para el primer Wittgenstein existe *el* lenguaje (en singular), con estructura coincidente con la de la realidad, y más allá del lenguaje está lo inefable, en cambio para el segundo Wittgenstein existen tantos lenguajes como usos lingüísticos y las cuestiones filosóficas (antes más allá del lenguaje) deben ser objeto de descripción superadora del embrujo del lenguaje. Mientras el primer Wittgenstein influyó en el positivismo lógico, en cambio el segundo influyó en la filosofía analítica, en especial en el llamado grupo de Oxford (Austin, Ryle, Strawson,etc.).

Pues bien, el segundo Wittgenstein dedicó sus esfuerzos intelectuales fundamentalmente a la filosofía de la matemática y a la filosofía de la mente. El mismo, al final de *Philosophical Investigations* (obra en la que centraremos nuestro estudio), señala un paralelismo entre ambas disciplinas filosóficas, al decir que es posible una investigación en conexión con las matemáticas que es enteramente análoga a su investigación sobre la psicología, que tal investigación es tan poco una investigación matemática como la segunda es una investigación psicológica, y que podría merecer el nombre de investigación sobre los fundamentos de la matemática. Es decir, podemos comentar, que para Wittgenstein la filosofía de la mente no es psicología científica sino fundamentación filosófica de la psicología.

La filosofía del segundo Wittgenstein posee dos características externas que dan razón de algunas peculiaridades de sus doctrinas. En primer lugar, una extraordinaria honradez intelectual que le lleva a crear una nueva filosofía, sin pretender mantener o salvar las ideas del *Tractatus*, así como a no escamotear temas o aspectos en el análisis de las expresiones y conceptos psicológicos, por incómodos que sean. En el prefacio de las *Philosophical Investigations*, escrito en 1945, nuestro autor nos dice que desde que ha comenzado de nuevo a ocuparse en la filosofía se ha visto forzado a reconocer graves errores en lo que escribió en su primer libro. A su vez, la otra característica externa de la filosofía del segundo Wittgenstein es su constante duda acerca de las soluciones que encuentra, lo cual da razón de lo mucho que escribió sin llegar a publicarlo

y, asimismo, de las vacilaciones y estilo adogmático de sus *Philosophical Investigations*, que como ya hemos dicho él mismo preparó para la publicación.

2. Filosofía de la mente y el argumento del lenguaje privado

Una tentación frecuente y fácil es estudiar las doctrinas de Wittgenstein en filosofía de la mente ciñéndose a su discusión contra la posibilidad de un lenguaje privado. Nosotros empezaremos por ahí, ya que al hilo del argumento de Wittgenstein contra el lenguaje privado encontraremos tesis de interés, pero también nos ocuparemos de otros argumentos y consideraciones.

Es característico del segundo Wittgenstein la atención que presta a lo que considera los diversos juegos lingüísticos realmente practicados por los seres humanos. Dicho rápidamente, un juego lingüístico (*language-game*, en alemán *Sprachspiel*) es un uso de un tipo de lenguaje. Nuestro autor acuñó este término por la comparación constante que es posible hacer, y él hace, entre juegos lingüísticos y juegos en general. Ejemplos de juegos lingüísticos, puestos por Wittgenstein, son: dar órdenes y obedecerlas, describir la apariencia de un objeto o dar sus medidas, construir un objeto a partir de una descripción (un dibujo), contar un suceso, especular acerca de un suceso, formar y comprobar una hipótesis, presentar los resultados de un experimento en cuadros y diagramas, inventar una historia y leerla, representar teatro, jugar al corro, adivinar acertijos, hacer una broma, contarla, resolver un problema en aritmética práctica, traducir de un lenguaje a otro, pedir, agradecer, maldecir, saludar, rezar. Puede advertirse que la noción de juego lingüístico es muy amplia, ya que incluye actividades específicamente basadas en palabras (orales o escritas), actividades donde las palabras son un componente más, e incluso actividades sin palabras que son expresivas de algún modo. Por tanto, el lenguaje se refiere en este contexto a cualquier tipo de signos, mientras que la noción de juego lingüístico se refiere a cualquier uso de un tipo de signos.

A su vez, cada juego lingüístico supone una determinada conducta o comportamiento de quien lo juega, puesto que, podemos decir, no nos comportamos de la misma manera al maldecir que al rezar, al ordenar que al obedecer, etcétera.

Los juegos lingüísticos interesan a Wittgenstein no en cuanto a la descripción estática de sus elementos sino en cuanto a la observación dinámica de sus reglas. Para él, hablamos acerca de los fenómenos del lenguaje espaciales y temporales, no acerca de un absurdo inespacial e intemporal, pero hablamos acerca de ello tal como lo hacemos acerca de las piezas del ajedrez, al indicar sus reglas de juego, no al describir sus propiedades físicas. Así pues, la observación del uso del lenguaje no es sino una observación de sus reglas. Y justo en este punto la noción fundamental para Wittgenstein es la noción de seguir una regla.

¿Qué es seguir una regla? En primer lugar, seguir una regla u obedecerla supone una obligación por parte del que la sigue así como un apremio o fuerza por parte de la regla, de modo que el que sigue la regla advierte que algo debe ser hecho. En segundo lugar, una regla no es una invitación a una acción aislada o a una persona determinada, sino que tiene un carácter más o menos general, siendo algo repetitivo y compartido socialmente. Y finalmente, una regla puede resultar dudosa en su aplicación, de tal manera que al seguir una regla podemos equivocarnos. Más profundamente, según Wittgenstein, una regla es una praxis social y pública. En efecto, en el parágrafo 202 de *Philosophical Investigations* nos dice que obedecer una regla es una práctica y que *creer* que uno está obedeciendo una regla no es (sin más, podemos añadir) obedecer una regla; por ello, continúa, no es posible obedecer una regla "privadamente", ya que de lo contrario creer que uno está obedeciendo una regla sería lo mismo que obedecerla. En suma, para seguir una regla no basta con creer que se la sigue, sino que es preciso seguirla realmente, y para ello debe ser seguida públicamente.

Con esta consideración se abre el debatido tema del rechazo por Wittgenstein de un lenguaje privado, así como de sus implicaciones. Al respecto sostendremos dos tesis: 1) el

carácter público de las reglas determina la propia publicidad del lenguaje, y 2) Wittgenstein admite experiencias privadas aunque no admite un lenguaje privado. En efecto, y respecto de la primera tesis, las reglas son, según Wittgenstein, instituciones, costumbres, en definitiva algo social y público, pero no algo individual y privado; ahora bien, como las reglas del lenguaje determinan el lenguaje o, más exactamente, cada juego lingüístico está caracterizado por ciertas reglas, tal carácter público de las reglas exige la publicidad del lenguaje; en consecuencia, no existe un lenguaje privado que merezca llamarse lenguaje propiamente dicho.

Respecto de la segunda tesis, nuestro autor admite el carácter privado de nuestras experiencias, de tal modo que, por ejemplo, nuestras sensaciones no son compartidas por los demás. Es indudable que si alguien golpea tanto a mí como a mi acompañante, mi dolor no es el suyo ni el suyo es el mío. Y sin embargo, cuando hablamos de dolor, al nombrarlo en un lenguaje, ninguno de ambos resulta privilegiado. Así pues, mientras las sensaciones, por ejemplo, son privadas, en cambio el lenguaje que las expresa es público y compartido. Para Wittgenstein, la propia palabra "sensación" es una palabra de nuestro lenguaje común, y no de un lenguaje privado que sólo yo entienda. En resumen, nuestra experiencia privada se expresa en un lenguaje necesariamente público.

Al hilo del argumento de Wittgenstein contra el lenguaje privado, parece claro que para nuestro filósofo los procesos mentales sólo pueden ser estudiados mediante su manifestación en un lenguaje público. Esto supone una actitud antimentalista que es general, como veremos, en Wittgenstein, aunque tal actitud es matizada en diversas ocasiones con reservas de tipo mentalista. Asimismo tal posición antimentalista supone favorecer una psicología conductista; en efecto, si los procesos mentales (y el análisis de Wittgenstein de las sensaciones puede generalizarse) sólo se conocen públicamente, es forzoso recurrir a la conducta, sea o no específicamente lingüística, para conocer tales procesos, pues sólo la conducta es pública. Como dice nuestro autor, si tengo que imaginar el dolor de otro según el modelo de mi propio dolor, entonces esto

no es una cosa tan fácil, ya que tengo que imaginar un dolor que no siento según el modelo de un dolor que siento. Es decir, no podría proyectar mi dolor sobre el dolor del otro, sino que tendría que atenerme a la conducta, que es pública.

3. Conductismo y mentalismo

De acuerdo con esta inclinación al conductismo, Wittgenstein muestra grandes reservas respecto de la introspección. En la segunda parte de las *Philosophical Investigations* dice tajantemente: "no intentes analizar tu propia experiencia interna". Y en el parágrafo 587 declara que a veces se podrá decir que sabemos algo por introspección, pero que en la mayoría de los casos no se podrá decir tal cosa. Así pues, podemos advertir que para Wittgenstein la introspección es un hecho real, pero no es fuente de auténtico conocimiento.

Nuestro filósofo analiza los fenómenos psicológicos de hablar para sí o internamente (sin proferir palabra alguna) así como los de los cálculos mentales. Tales fenómenos interesan en cuanto no suponen conducta visible alguna. Sin embargo, Wittgenstein considera que estos fenómenos especiales están supeditados respectivamente al habla externa y al cálculo perceptible. El psicólogo conductista Watson, en su libro de 1925, *Behaviorism,* había escrito que su teoría sostiene que los hábitos musculares aprendidos en el habla externa son responsables del habla interna o implícita (pensamiento). A su vez, nuestro autor señala que el íntimo parentesco entre el "habla interna" y el "habla" se manifiesta en la posibilidad de decir en voz alta lo que uno dijo internamente y en que al habla interna la puede acompañar una acción externa. Por otra parte, en el caso de los cálculos mentales, según Wittgenstein, sólo podemos aprender a calcular "en la cabeza" si aprendemos a calcular de modo perceptible. En suma, no se rechaza la legítima existencia de estos fenómenos, pero se insiste en la primacía de la conducta. O dicho de otra manera, Wittgenstein acepta la existencia de procesos mentales internos (mentalismo) pero sitúa en la conducta la medida objetiva de tales procesos.

Sin embargo, y para incomodidad del intérprete simplista, en el parágrafo 308, donde nuestro autor se plantea el problema de los procesos mentales en relación con el conductismo, se muestra inclinado a cierto grado de mentalismo que, no obstante, disolverá en gran medida en otros textos. En ese parágrafo advierte dos etapas o fases al plantearnos la relación entre procesos mentales y la conducta. En una primera etapa hablamos de los procesos y estados mentales dejando indeterminada su naturaleza, pensando que en algún momento podremos llegar a saber algo preciso sobre tal naturaleza. Esta primera etapa, podemos comentar, sería la propia de una psicología conductista no beligerante, es decir, de un conductismo metodológico: debemos atenernos a la conducta aunque sin negar procesos mentales al margen de la conducta. Pero para Wittgenstein tal primera etapa nos ata a una segunda etapa, donde los procesos mentales ya no son puestos entre paréntesis sino negados: por tanto hemos de negar el proceso aún incomprendido en el medio aún inexplorado. Nuestro filósofo compara esta situación a un truco de prestidigitación: primero parece que tenemos procesos mentales y a continuación éstos desaparecen. Podríamos decir que de la simple supresión metodológica hemos pasado a la supresión ontológica. Y sin embargo, el parágrafo 308 termina diciendo que no queremos negar tales procesos espirituales (*geistige Vorgänge*), con lo que tenemos aquí una clara apertura al mentalismo. En suma, podemos añadir, Wittgenstein considera natural el paso al conductismo, aunque los procesos mentales, incluso desconocidos, permanecen; según ello, el conductismo sería la manera normal de considerar las cosas en psicología, pero no podemos eliminar los procesos mentales (mentalismo).

Con todo, esta concesión al mentalismo parece ser para nuestro filósofo tan necesaria como inútil, con lo que su posición básica seguirá siendo antimentalista. En efecto, y en un parágrafo algo anterior, donde se refiere al proceso psicológico de recordar, admite que hay un proceso interno pero añade que tal figura de un proceso interno resulta inútil para el empleo de la palabra "recordar". Wittgenstein sabe que

podemos fingir y que, en esa medida, la conducta no es enteramente fiable, y sabe también que hay procesos mentales distinguibles de la conducta, pero con todo y de nuevo, para él, el proceso mental interno no presta servicio alguno y es como nada. Por supuesto, señala, hay una diferencia entre conducta de dolor con dolor y conducta de dolor sin dolor, pero acerca de la sensación (interna) nada puede decirse. Esta idea de que el proceso mental no se puede decir o describir aparece también en el parágrafo 662: uno puede decir que las palabras "yo quisiera que fulanito viniera a verme" describen un estado mental, pero eso no se puede decir.

Las reticencias de Wittgenstein respecto del mentalismo son más evidentes cuando dice que los procesos mentales, en cuanto internos o ajenos a la conducta, son curiosos o extraños, que su naturaleza es comparable a la del tiempo y que, como de éste, hablamos de procesos mentales sin saber a ciencia cierta de qué hablamos.

Pero aunque Wittgenstein es ante todo un autor antimentalista, aunque rechaza la utilidad para la psicología científica de recurrir a nociones puramente mentalistas, con todo (y para desesperación de dogmáticos) el ingrediente mentalista aparece reiteradamente en sus textos, a pesar de sus dificultades. Un nuevo ejemplo: en el parágrafo 421 señala que nos parece paradójico que, en un mismo informe, mezclemos los estados del cuerpo y los estados de la conciencia, ya que lo primero es tangible mientras que lo segundo es intangible y, sin embargo, lo intangible también proporciona un uso.

Podría pensarse que las vacilaciones de Wittgenstein se resolverían si hubiese optado por una teoría de la identidad entre cerebro y mente en términos tales que los procesos mentales internos, no eliminables, se reducirían a procesos cerebrales, que serían a su vez irrelevantes para la explicación psicológica, ya que ésta radicaría en un mero análisis conductual. Pero, aparte de que el problema en las *Philosophical Investigations* no es la relación mente-cerebro sino la relación mente-conducta, hay textos de Wittgenstein que rechazan la identificación entre cerebro y mente. En el parágrafo 427 nos dice que cuando hablamos de lo que ocurre

"dentro de la cabeza" no se piensa en procesos cerebrales sino en procesos de pensamiento. A su vez, en *Zettel*, obra póstuma editada en 1967, en los parágrafos 605 y siguiente, Wittgenstein sostiene que una de las ideas más peligrosas para un filósofo es que pensamos con nuestras cabezas o en nuestras cabezas, añadiendo que la idea del pensamiento como un proceso en la cabeza, en un espacio completamente cerrado, nos da algo oculto. En suma, los procesos mentales no se encierran en el cerebro, de tal manera que el mentalismo problemático e inseguro de nuestro autor permanece como tal.

4. La tarea de la psicología

Dados estos presupuestos, ¿cuál es la tarea de la psicología, cómo debe proceder? En la sección V de la segunda parte de las *Philosophical Investigations* Wittgenstein escribe sobre este asunto oscilando como de costumbre entre un antimentalismo básico y un mentalismo residual. En efecto, la psicología para él se ocupa de la conducta pero pretende saber de la mente a través de la conducta; y por otra parte la mente, en cuanto distinta de la conducta, constituye una presuposición tácita. Los psicólogos, viene a decirnos, registran y observan la conducta de los seres humanos, pero tales observaciones conductuales pueden estar al servicio del conocimiento de los estados mentales expresados, y entonces la mente aparece como un supuesto implícito. Así pues, en psicología hay dos distintos juegos lingüísticos: uno es la "expresión" de estados anímicos, otro es la "comunicación" de conducta.

Estas consideraciones pueden aclararse recurriendo al parágrafo 472 de *Zettel*, donde Wittgenstein establece un plan para el tratamiento de los conceptos psicológicos. Aquí distingue entre las sentencias en tercera persona del presente, que dan información o comunicación, y las sentencias en primera persona del presente, que proporcionan expresión; aunque nuestro autor advierte que la distinción no es completamente exacta. Podemos pensar que tal distinción no es enteramente exacta porque los informes en tercera persona pueden incluir

expresiones, como al decir "él se queja de dolor" puedo incluir la expresión de mi simpatía hacia el otro hasta pensar que se queja de dolor cuando sólo está inquieto; a su vez, una expresión en primera persona puede incluir comunicación, como al decir "yo tengo dolor" si no finjo. En todo caso, resulta claro que una psicología conductista es una psicología de informes o comunicaciones en tercera persona: "él hace cálculos", "él se ríe", "él se agita de dolor", etc. En cambio, una psicología no-conductista es una psicología de expresiones en primera persona: "yo hago cálculos", "yo me río", "yo me agito de dolor", etcétera.

En la sección XI de la segunda parte de las *Philosophical Investigations* varios textos apuntan la idea de que para Wittgenstein saber o conocer quiere decir adquirir evidencia o recibir información, que en cuanto tales son falibles. En efecto, alguien no "sabe" algo cuando la duda que existe para mí no existe para él; o también, el saber no excluye lógicamente la duda, de tal manera que hablamos de saber cuando podemos convencernos de algo. De modo aún más claro, nuestro autor dice, en la misma sección, que yo puedo saber lo que el otro piensa, pero no puedo saber lo que yo pienso, siendo correcto decir "sé lo que piensas" y siendo falso decir "sé lo que pienso". Esto es, podemos comentar, el saber acerca de mis pensamientos tiene una certeza que hace que no sea propiamente saber o conocer. En estas circunstancias, ya que sólo hay saber o conocimiento cuando la duda es posible, sólo habrá conocimiento o saber en una psicología conductista de tercera persona. O dicho de otro modo, las certezas que yo puedo tener acerca de mis estados mentales no son ciencia psicológica, las sentencias en primera persona no son propias de la psicología; en cambio, los conocimientos (susceptibles de duda) acerca de la conducta de otros son ciencia psicológica, las sentencias en tercera persona son propias de la psicología.

El antimentalismo de Wittgenstein, que con todo admite la existencia de la mente como presupuesto implícito, también se advierte al señalar la necesidad de criterios externos para poder llegar a conocer los procesos mentales. En el parágrafo 580 de *Philosophical Investigations* afirma tajantemente que

un "proceso interno" necesita criterios externos. Los procesos mentales pueden disponer de criterios para su aclaración ante el sujeto, criterios que serán privados e internos, pero para su conocimiento en sentido estricto, para su saber objetivo, es preciso criterios externos apreciables en la conducta. Wittgenstein nos dice que no nos interesa nada de lo que ocurre dentro del jugador de ajedrez cuando queremos saber si sabe jugar al ajedrez, sino que nos interesan los criterios (externos) que nos demuestran su capacidad, y que tales criterios son distintos de los criterios de los "estados internos". Más aún, los criterios externos son los modelos de esos posibles criterios internos, al igual que, como vimos anteriormente, el cálculo perceptible es el modelo del cálculo mental.

En resumen, Wittgenstein simpatiza con una psicología científica básicamente conductista, centrada en el análisis de la conducta, lingüística o no, en cuanto algo público y objetivo. Y sin embargo, no comparte la eliminación de la "vida mental" habitual entre los psicólogos conductistas. Tal vida mental, constituida por procesos internos, no es identificada con los procesos cerebrales, permaneciendo como algo tan nebuloso como recalcitrante a su desaparición. Antimentalismo y mentalismo coexisten, a pesar de las dificultades, en la filosofía de la mente del dubitativo y adogmático filósofo vienés.

5

Crítica del conductismo

1. Conductismo metodológico y conductismo ontológico

Hemos considerado hasta aquí, entre otras cosas, la filosofía de la mente de Gilbert Ryle y de Ludwig Wittgenstein, filósofos ambos inclinados a la psicología conductista. Corresponde ahora examinar el conductismo entre los propios psicólogos.

Podemos comenzar señalando qué es el conductismo en general y distinguir luego entre conductismo metodológico y conductismo ontológico. En general, el conductismo es la doctrina que sostiene que la psicología debe limitarse al estudio de la conducta, tanto en los animales como en las personas, es decir, debe limitarse al estudio de la actividad de los sujetos que sea públicamente observable y registrable, incluyendo en tal actividad los movimientos corporales, ruidos producidos, substancias químicas emitidas, interacciones con el ambiente, etc. En consecuencia, debemos rechazar cualquier aproximación a la psicología de corte dualista o de sesgo introspectivo, de tal manera que no es aceptable introducir distinción alguna entre cuerpo y mente, ni tampoco recurrir a la introspección (o examen por el propio sujeto de sus procesos mentales) como método de la psicología. El conductismo concibe la psicología como una ciencia natural, análoga a la física o a la química, y entiende que el único ámbito de su estudio es el mundo material, tal como ocurre en

estas ciencias. La introspección, en cuanto algo privado e individual, no es fiable cuando, como sucede en el conductismo, lo que importa es la observación pública e intersubjetiva.

Ahora bien, cabe distinguir dos formas posibles de conductismo. Por un lado, lo que podemos denominar conductismo metodológico, esto es, la doctrina según la cual la limitación de la psicología al estudio de la conducta está justificada porque la investigación acerca de cualquier otro aspecto de los sujetos no puede ser comprobada de acuerdo con un método científico; para el conductismo metodológico la limitación al estudio de la conducta es consecuencia del hecho de que sólo la conducta puede someterse al método científico. En este punto debe señalarse que, salvo algún autor aislado o algunas declaraciones no sistemáticas, la mayoría de los psicólogos conductistas practican el conductismo metodológico, es decir, se limitan al estudio de la conducta ignorando cualquier otro aspecto por razones de (supuesta) eficacia metodológica. Por otro lado, cabe hablar de conductismo ontológico, esto es, la doctrina según la cual la limitación de la psicología al estudio de la conducta está justificada plenamente porque lisa y llanamente no existen procesos mentales ajenos a la conducta, es decir, la única realidad que existe para el psicólogo es la conducta; en este caso, pues, sólo tiene sentido ocuparse de procesos conductuales y no tiene sentido ocuparse de procesos mentales porque la mente como tal no existe.

2. Breve presentación del conductismo

El fundador de la psicología conductista fue el americano John Watson (1878-1958). Habiéndose doctorado en psicología en la Universidad de Chicago en 1903, Watson fue nombrado ayudante en esta Universidad. Admirador de los trabajos de Edward Thorndike (1874-1949) acerca de la inteligencia animal, desarrolló amplias observaciones de monos, ratas blancas y pájaros para intentar comprender su conducta. De este modo, desde sus inicios Watson asoció estrechamente la investigación de la conducta animal con la investigación de la conducta humana.

En 1908 Watson fue nombrado catedrático de psicología en la Universidad Johns Hopkins, de Baltimore, donde estableció un laboratorio de investigación en psicología animal. A esta época pertenece su célebre artículo "Psychology as a Behaviorist Views it" (1913), que constituye el primer manifiesto de la psicología conductista. La tesis central de este trabajo es que la psicología debe ser la ciencia de la conducta humana, la cual, al igual que la conducta animal, tiene que ser estudiada en las condiciones de precisión y exactitud propias del análisis de laboratorio.

Watson publicó otros trabajos que alcanzaron gran éxito e incluso divulgación popular. Citemos, a título de ejemplos, su libro *Behavior: An Introduction to Comparative Psychology* (1914) y *Psychology from the Standpoint of a Behaviorist* (1919), un libro donde establece sus tesis definitivas. En 1921, tras dejar la Universidad Johns Hopkins, empezó su carrera en el mundo del negocio publicitario, del cual se retiró en 1946. Entretanto, en 1925, publicó su libro *Behaviorism*, destinado al gran público y en donde expone en términos de divulgación sus ideas acerca del conductismo.

Sería muy largo y tarea fuera de lugar exponer de modo completo la psicología conductista de Watson, razones por las cuales me limitaré a resumir sus principales puntos de vista en relación con el conductismo. Ya en su citado artículo de 1913, Watson establece que la finalidad de la psicología conductista es la predicción de la conducta, humana y animal, así como su control. Precisamente esta idea del control de la conducta, presente en la mayoría de los psicólogos conductistas, implica un programa ambicioso de control del hombre por la ciencia, con consecuencias evidentes actualmente en la tecnología publicitaria. Asimismo nuestro autor rechaza enérgicamente la introspección, declarando en el citado artículo que la observación introspectiva no forma parte esencial de los métodos de la psicología. Finalmente, en el mismo lugar, Watson mantiene que se debe excluir en una psicología científica cualquier referencia a la conciencia y a sus estados, lo cual hará posible eliminar la barrera entre la psicología y la ciencia natural.

Para Watson la cadena estímulo-respuesta explica cualquier fenómeno psicológico, tanto de los seres humanos como de los animales. Y, en cuanto al método, el psicólogo no debe estudiar su propia conducta sino la de los otros. En concreto, el objetivo de los estudios psicológicos es encontrar los datos y leyes mediante los cuales, dado el estímulo, se puede predecir la respuesta, o bien, dada la respuesta, se puede determinar la naturaleza del estímulo. En consecuencia, el estímulo aparece como la causa de la respuesta, sin referencia a factor interno alguno.

Así pues, el par estímulo-respuesta constituye el resorte explicativo básico de los fenómenos psicológicos para el conductismo, de tal manera que el estudio de la conducta consiste en establecer relaciones entre estímulos y respuestas en cualquier organismo. Ahora bien, estímulo es el conjunto de las excitaciones que actúan sobre el organismo en un momento dado, y a su vez el ambiente del organismo constituye un complejo de estímulos para él, complejo que determina la conducta. En este punto podemos observar que la conducta es básicamente un conjunto de respuestas, de tal modo que, aunque el esquema concreto conductista es el par estímulo-respuesta, el esquema general es el par ambiente-conducta. Sin embargo, y como primera observación crítica, frecuentemente es difícil decir si una determinada acción (tal como una secreción glandular) es estímulo o respuesta. A su vez, respuesta es el conjunto de cambios que se producen en los músculos y en las secreciones glandulares.

Por otra parte, una característica notable del conductismo es la exclusión de las explicaciones neurológicas. Watson llega a decir que es perfectamente posible estudiar la conducta sin saber nada acerca del sistema nervioso periférico e incluso acerca del sistema nervioso central. Con ello no sólo rechazaba el recurso a la conciencia propio de la psicología tradicional sino también el recurso a la neurología. En consecuencia, el ser humano que estudia la psicología conductista está enteramente desposeído de estados internos, ni mentales en el sentido tradicional ni tampoco estados cerebrales.

El otro gran representante de la psicología conductista es

Burrhus Skinner (1904-1990), a quien su larga vida y su constante actividad de investigador y escritor permitieron ejercer una gran influencia sobre la psicología del siglo xx. Su primera gran obra fue publicada en 1938 bajo el título *The Behavior of Organisms*, en la que asumía el conductismo de su compatriota Watson. Durante la Segunda Guerra Mundial entrenó palomas para pilotar torpedos y bombas, aunque nunca fueron utilizadas. Su fe en la posibilidad de desarrollar una tecnología de la conducta, tanto en humanos como en animales, le llevó a sobrevalorar la capacidad de entrenamiento de los animales. Esta misma fe en la tecnología de la conducta se manifiesta en su popular novela *Walden Two* (1948), donde se presenta una comunidad utópica organizada y desarrollada según los principios de la tecnología conductista.

En 1948, Skinner fue nombrado catedrático de psicología en la Universidad de Harvard, donde permaneció hasta su jubilación en 1972. Aquí escribió su obra fundamental *Science and Human Behavior* (1953), seguida algunos años más tarde de *Verbal Behavior* (1957). Skinner incorporó a la psicología conductista la noción de conducta operante (contrapuesta a lo que podemos denominar conducta pasiva); con ella se refiere a la conducta que depende más del refuerzo o recompensa, tras el éxito alcanzado, que de los estímulos espontáneos; en efecto, el refuerzo aumenta la probabilidad de la conducta en cuestión. En este punto cabe observar que tenemos un caso de respuesta activa, no meramente pasiva, ya que el sujeto modifica su conducta en función del éxito alcanzado

En cuanto a la posición general conductista, mientras Watson mantuvo básicamente un conductismo metodológico con algunas declaraciones a favor del conductismo ontológico, en cambio Skinner, en su larga actividad científica, sostuvo diversas posturas. En "Behaviorism at fifty" (1964) defendió el conductismo ontológico, es decir, la negación explícita de la mente como algo distinto de la conducta. A su vez, en su polémico libro *Beyond Freedom and Dignity* (1971), donde considera que las nociones de libertad y dignidad pueden conducir a los seres humanos a la autodestrucción, además de insistir en el papel del ambiente como determinante de las

peculiaridades de las personas, Skinner sostiene que el campo de la conciencia existe realmente aunque quede fuera de la consideración científica por su misma naturaleza. Finalmente, en *About Behaviorism* (1974) nuestro autor propone una versión del conductismo en la que se asevera la realidad de los fenómenos internos e incluso el acceso introspectivo a ellos, aunque tal realidad de los fenómenos mentales no es no-física sino fisiológica y además la introspección tiene un carácter limitado y poco fiable. En suma, lo que permanece constante en los escritos de Skinner es la importancia decisiva otorgada al ambiente en la modelación de la conducta, así como su fe en la posibilidad de desarrollar una eficaz y benéfica tecnología de la conducta que produzca mejores seres humanos.

3. Valoración crítica del conductismo

Aunque la psicología conductista está aún viva y presente entre numerosos investigadores, ha perdido su hegemonía desde más o menos 1960 en favor de la psicología cognitiva (de la cual nos ocuparemos en detalle más adelante), la cual se caracteriza básicamente por postular en cada sujeto un sistema de estados mentales internos gobernado por procedimientos computacionales. Estos dos elementos, mentalismo y tratamiento computacional, son completamente ajenos al conductismo. Sin embargo, no ha sido solamente la fuerza de la psicología cognitiva la razón del progresivo desplazamiento de la psicología conductista sino de modo aún más definitivo las debilidades del conductismo que pasamos a analizar.

Para el conductismo cualquier tipo de fenómeno psicológico (como una sensación, un sentimiento, una imagen mental o un pensamiento) se identifica, o al menos se asocia estrechamente, con un tipo determinado de conducta. Pero tal tesis, sostenida de modo general, es falsa. Para que fuese verdadera debería ocurrir siempre que cualquier fenómeno psicológico estuviese asociado o identificado con cierta conducta específica, y asimismo que cualquier conducta estuviese asociada o identificada con cierto fenómeno psicológico también específico. Sin embargo, existen fenómenos psicológicos a los que no

corresponde conducta alguna (en cuanto actividad públicamente observable y registrable), y asimismo existen conductas que no corresponden a un fenómeno psicológico determinado de modo inequívoco. Véamoslo por separado.

En efecto, y en primer lugar, un sujeto puede tener un fuerte dolor de cabeza y no manifestarlo, o bien sentir un cosquilleo en la espalda y prescindir de rascarse. Considerando fenómenos psicológicos más elaborados, una persona puede experimentar en un momento determinado un sentimiento de odio o de envidia hacia otra persona y ocultar tales sentimientos, de tal modo que no se manifiesten en su conducta; o bien, un hombre puede imaginar que besa apasionadamente a la mujer con la que está hablando y tal imagen mental pasa enteramente desapercibida para aquéllos que le observan. Más aún, cuando comprendemos una idea o una teoría tal comprensión no se manifiesta necesariamente en conducta alguna, de tal manera que, por ejemplo, a un profesor le resulta imposible saber, sin interrogatorio adicional, si determinado alumno ha entendido o no la lección impartida. La razón de todo ello es que sensaciones, sentimientos, imágenes mentales, pensamientos y, en general, los fenómenos psicológicos, son, contrariamente a lo que sostiene el conductismo, procesos mentales internos, que pueden reflejarse o no en la conducta, pero que no están necesariamente asociados ni tampoco identificados con una conducta determinada. En suma, y frente al conductismo, son posibles estados mentales sin conducta.

En segundo lugar, un sujeto puede decirnos que le duele la cabeza de un modo terrible, o incluso quejarse y llevarse las manos a la cabeza mostrando una conducta de dolor de cabeza, y sin embargo no tener dolor de cabeza alguno. O también puede rascarse frenéticamente la espalda sin sentir cosquilleo alguno en tal lugar. En el caso de fenómenos psicológicos más elaborados, una persona puede comportarse respecto de otra dando muestras de amistad o admiración sin que realmente tenga tales sentimientos hacia ella; es más puede ocurrir que realmente la odie y la desprecie. Asimismo una persona puede declarar que recuerda la conversación

mantenida con otra, cuando en realidad no existe en ella tal recuerdo. Finalmente, un sujeto puede decir que ha entendido una teoría o una idea y, sin embargo, no existe en su mente la comprensión de tal teoría o idea. Todos estos ejemplos son casos de fingimiento de conducta y señalan claramente que hay conductas que no corresponden a los fenómenos psicológicos a los cuales habitualmente se asocian. Una vez más se pone de relieve que los fenómenos psicológicos son procesos mentales internos, que pueden, en estos casos, estar ausentes a pesar de estar presentes las conductas habitualmente asociadas. En suma, y frente al conductismo, son posibles conductas sin los correspondientes estados mentales.

En consecuencia, puesto que son posibles estados mentales sin conducta y asimismo conductas sin estados mentales, no existe ninguna identidad o asociación estrecha entre estados mentales y conducta.

Watson y otros psicólogos conductistas han intentado explicar el fenómeno psicológico del pensamiento, al cual no corresponde en numerosas ocasiones una conducta (en cuanto actividad públicamente observable y registrable), señalando que se produce una conducta especial consistente en un lenguaje subvocal, con ligeros movimientos de los labios, de la garganta y de las cuerdas vocales. Sin embargo, la investigación experimental demuestra que si a un sujeto se le paralizan los labios, garganta y cuerdas vocales, con todo, su pensamiento continúa. Ello indica una vez más que el pensamiento es un proceso mental interno no reducible a conducta.

El conductismo está expuesto además a otras objeciones básicas que lo refutan. Para empezar, la psicología conductista, en buena parte de sus autores, supone una decidida actitud ambientalista que resulta insostenible. Por actitud ambientalista se entiende el apoyo a la tesis de que la conducta del sujeto se determina exclusivamente, o al menos de forma muy determinante, por las circunstancias de su ambiente. Con ello se niega o se minimiza la influencia de los aspectos hereditarios en la conducta. Skinner, en *Beyond Freedom and Dignity*, insiste en este punto defendiendo que las personas son extraordinariamente diferentes en diferentes lugares justa-

mente por causa de los lugares; más aún, para Skinner, el ambiente reemplaza la función y el papel del hombre autónomo. Asimismo la actitud ambientalista elimina cualquier competencia que pueda considerarse innata. Precisamente Noam Chomsky, en "A Review of B. F. Skinner's *Verbal Behavior*" (1959), sostiene que la predicción de la conducta de un organismo complejo requiere, además de la información acerca de la estimulación externa (a la que se limita Skinner), un conocimiento de la estructura interna del organismo, así como de los modos en los que tal organismo procesa la información de entrada y organiza su propia conducta. Tales características del organismo son, señala Chomsky, un complicado producto de una estructura innata.

Existe evidencia experimental en contra de la tesis defendida por muchos psicólogos conductistas de que la conducta es función exclusiva de la estimulación externa. Por ejemplo Karl Lashley (1890-1958), discípulo inicialmente de Watson y autor de *Brain Mechanisms and Intelligence* (1929), ha probado mediante experimentos que la eficacia de un estímulo depende de manera decisiva de la preparación o disposición del sujeto, la cual no siempre es controlable por el psicólogo.

Por otra parte, el ambientalismo supone lo que Mario Bunge y Rubén Ardila, en su libro *Filosofía de la Psicología* (1988), han denominado caja-negrismo, es decir, la estrategia de modelar sistemas como cajas vacías que responden únicamente a estímulos del medio. La psicología que resulta de la adopción de la estrategia de la caja negra, señalan Bunge y Ardila, se denomina psicología de estímulo-respuesta (E-R), mientras que si también toma en cuenta los estados internos del organismo tenemos una psicología de estímulo-organismo-respuesta (E-O-R), y si además explica las consecuencias que la respuesta tiene sobre la modalidad en que el sujeto elabora el input o entrada (esto es, si se toma en cuenta los giros de la retroalimentación), tenemos una psicología de estímulo-organismo-respuesta-retroalimentación (E-O-R-R). Muchos psicólogos conductistas adoptaron el caja-negrismo, es decir, una simple psicología de estímulo-respuesta, pero esto constituye un grave error. En efecto, la respuesta de

cualquier sujeto es una función no sólo del estímulo sino también del estado interno del sujeto. Más aún, podemos añadir, la psicología de simple estímulo-respuesta no proporciona explicaciones auténticas, ya que no explicita un mecanismo explicativo que medie entre estímulo y respuesta, sino que solamente describe una supuesta causa para un supuesto efecto.

Algunos psicólogos conductistas, conscientes de las deficiencias del caja-negrismo, decidieron introducir variables intermediarias entre estímulo y respuesta, construyendo así una psicología de estímulo-organismo-respuesta. Surgió de este modo el llamado neoconductismo, representado, entre otros, por Clark Hull (1884-1952), Edward Tolman(1886-1959) y Kenneth Spence (1907-1967). Sin embargo, el neoconductismo llevaba en sí mismo la ruina del conductismo. En efecto, y por un lado, la introducción de variables intermediarias carecía en algunos casos de un criterio, de tal modo que recurrir o no a ellas se convertía en asunto de pura convención. Y por otro lado, tal introducción de variables intermediarias aproximaba inevitablemente el conductismo al mentalismo, su enemigo inicial, ya que algunas variables se entendían como impulsos, propósitos o expectativas. Por ejemplo, Tolman, en *Purposive Behavior in Animals and Men* (1932), sostenía la idea de conducta dotada de propósitos y que busca la obtención de una meta determinada; pero tal idea general aparece en psicólogos cognitivos, de corte mentalista, como Miller, Galanter y Pribram en su libro conjunto *Plans and the Structure of Behavior* (1960), obra que constituyó un auténtico manifiesto de la psicología cognitiva.

6

Mentalismo y psicología popular

1. Mentalismo frente a conductismo

El mentalismo es una doctrina, popular, filosófica y científica, que insiste en el carácter interno y privado de los procesos mentales. En cuanto internos, el mentalismo considera que los procesos mentales no son conducta públicamente observable y registrable, sino o bien ajenos a la conducta o bien causas de la conducta (y como tales distintos de ella). Además, en principio, este carácter interno puede estar determinado por la índole cerebral del proceso mental, pero también puede venir determinado por la índole no-física del proceso mental; por ello, en principio, "interno" es un término neutral, en el sentido de que puede querer decir cerebral pero no necesariamente.

A su vez, en cuanto privados, el mentalismo sostiene que los procesos mentales de cada individuo no son compartidos por los demás (y así, por ejemplo, *mi* pena no es *tu* pena), y sobre todo sólo el propio individuo tiene acceso directo a sus procesos mentales, debiendo recurrir los demás individuos a vias indirectas para el conocimiento de tales procesos, ya sea el testimonio del sujeto en cuestión ya sea el conocimiento por analogía. Este privilegio del acceso directo de cada individuo a sus propios procesos mentales puede ir unido al carácter incorregible de tal conocimiento directo. Con ello se quiere indicar que nadie puede corregir o desmentir a un sujeto respecto de lo que establece acerca de sus propios procesos

mentales; por ejemplo, nadie puede decirme que yo no siento pena si en realidad siento pena. Sin embargo, tal carácter incorregible de nuestro conocimiento directo de nuestros procesos mentales no debe exagerarse ni pretender su validez universal, puesto que es posible equivocarnos en la evaluación de nuestros propios procesos mentales y además la psicología científica puede ayudarnos, y normalmente nos ayuda, a conocer mejor lo que para nosotros son datos inmediatos.

De lo expuesto resulta claro que hay una oposición radical entre mentalismo y conductismo, ya que para el conductismo no existen procesos mentales ni internos ni privados. Por ello es presumible que el declive del conductismo implique un auge del mentalismo y, en particular, que dado el desprestigio general actual del conductismo se dé un paralelo prestigio del mentalismo. Tal mentalismo puede apreciarse en la llamada psicología popular, que será objeto de análisis y discusión en este capítulo.

2. Caracteres de la psicología popular

La psicología popular (*Folk Psychology*), que también puede denominarse psicología del sentido común o bien psicología ordinaria, es el conjunto de conocimientos que todos los seres humanos poseemos *de facto* acerca de los procesos mentales en sí mismos y en su relación con la conducta, en cuanto tal conjunto no procede de una enseñanza de la psicología científica sino de nuestras experiencias personales y de una transmisión informal de conocimiento por parte de las personas que a lo largo de nuestras vidas nos rodean.

La psicología popular constituye al mismo tiempo un marco conceptual y una serie de generalizaciones explicativas que emplean los conceptos de tal marco. En este marco se encuentran las nociones de creencia, deseo, temor, etc. A su vez, las generalizaciones explicativas son de los tipos "una persona no hará algo que cree que la perjudica", "una persona se empeñará en hacer algo que desea fervientemente", "una persona que teme a otra evitará tratarla", etc. Cada una de

ellas es utilizada espontáneamente para explicar la conducta de los seres humanos o sus estados de ánimo. Por ejemplo, y respectivamente, "Juan no votó a los socialistas porque cree que subirán los impuestos", "Antonio invita todos los fines de semana a Ana porque desea ser su novio", "Paco no asistió a la fiesta porque tiene miedo de encontrarse con Pepe".

Este acervo de conceptos y generalizaciones es común a todos los seres humanos, aunque no es exactamente igual en todos ni tampoco es infalible. En efecto, la psicología popular ha cambiado a lo largo de los tiempos, aunque sus nociones y resortes básicos sigan inalterables, incorporando o abandonando algunos elementos que tienen que ver con la popularización de la propia psicología científica. Asimismo tampoco es la misma en una u otra cultura, en función de la incorporación o no de los aspectos popularizados de la psicología científica y en función de las características de las mentalidades de los pueblos. Por ejemplo, un español dirá que "Pepe es agresivo porque tiene complejo de inferioridad", pero es muy improbable este modo de explicación en un pigmeo africano. Por otra parte, es indudable que ciertas personas poseen más *talento natural* que otras para entender los procesos mentales y la conducta del prójimo, y por ello decimos que fulanito es un buen psicólogo mientras menganito no es un buen psicólogo (sin pretender valorar sus conocimientos científicos).

Más importante es el hecho de que las generalizaciones de la psicología popular no pretenden ser, ni tampoco son, leyes estrictas, sino únicamente mecanismos útiles de explicación y de predicción en casos sencillos y cotidianos, y que incluso con estas limitaciones con frecuencia fallan. Ninguna persona razonable pretende explicar con sus dotes meramente naturales comportamientos complejos o estados de ánimo de personalidades extraordinarias. Puedo prever que Antonio me ayudará porque tiene sentimientos de amistad hacia mí, pero no pretendo evaluar con exactitud el alcance y duración de su ayuda.

La psicología popular puede compararse con la lógica natural y con la física popular. Dicho rápidamente, la lógica natural es el conocimiento que todos tenemos acerca de las

formas de argumentar y que no procede del estudio de la lógica científica, sino que poseemos de modo innato y es desarrollado junto con el aprendizaje del lenguaje. A su vez, la física popular es el conocimiento y habilidades que todos poseemos para movernos en nuestro entorno físico con éxito, y que no debemos al estudio de la física científica. No necesitamos estudiar lógica para argumentar que "si bebo mucho vino entonces me emborracho, y si estoy borracho entonces no conduzco bien, con lo que si bebo mucho vino no podré conducir bien". Tampoco necesito estudiar física para saber que no puedo pasar a través de una pared. Aunque la psicología popular, la lógica natural y la física popular atienden a necesidades distintas del ser humano, todas ellas coinciden precisamente en eso: son conocimientos ajustados a nuestras necesidades, presumiblemente presentes en los seres humanos desde el *homo sapiens*, y una conquista más de la evolución.

Ahora bien, la psicología popular va unida de modo muy estrecho a una actitud mentalista. Justamente por ello los psicólogos conductistas despreciaban la psicología popular y sostenían que las nociones de deseo, creencia, intención, sentimiento y, en general, cualquier noción que pretenda denotar un proceso mental interno deben ser eliminadas de la psicología científica. Como ya indicamos, el conductismo perdió su hegemonía en torno a 1960, siendo sustituido por la psicología cognitiva, la cual adopta una actitud mentalista frente al conductismo. Esta actitud mentalista se manifiesta asimismo en la psicología popular, ya que las creencias, sentimientos, deseos, temores, etc. se consideran popularmente como procesos mentales que son internos y privados. Popularmente entendemos que nuestros procesos mentales son episodios de nuestra *vida interior*, que sólo nosotros normalmente sabemos bien lo que nos pasa y que los demás han de limitarse a barruntarlo.

3. Jerry Fodor y la psicología popular

Puesto que la psicología cognitiva y la psicología popular coinciden en la actitud mentalista, no resultará extraño

encontrar firmes defensores de la psicología popular entre los psicólogos cognitivos. El autor más notable al respecto, al tiempo que uno de los máximos representantes de la psicología cognitiva, es Jerry Fodor. Sus puntos de vista sobre el valor de la psicología popular y sobre la relación entre psicología popular y psicología cognitiva se exponen nítidamente en su libro *Psychosemantics* (1987). Fodor se refiere a la psicología popular llamándola psicología del sentido común o, también, psicología de creencia/deseo del sentido común en atención al papel que estas nociones desempeñan en las explicaciones psicológicas del sentido común.

Para Fodor, la psicología popular está muy próxima a la verdad, ya que explica muchas cosas acerca de la conducta y su adecuación predictiva está más allá de la disputa racional. Por otra parte, la psicología popular resulta indispensable, puesto que no existe alternativa al vocabulario de la explicación psicológica del sentido común. Pero asimismo la psicología popular, según Fodor, es una psicología de actitudes proposicionales, ya que creencia y deseo son actitudes proposicionales. (Aclararé brevemente que una creencia o un deseo se consideran por muchos autores como actitudes proposicionales porque se supone que cuando un individuo cree o desea algo mantiene una actitud respecto de una proposición que expresa lo creído o lo deseado.)

Ahora bien, siempre según Fodor, las actitudes proposicionales se caracterizan de modo esencial por satisfacer las siguientes condiciones: 1) son semánticamente evaluables, es decir, sus contenidos poseen valores de verdad, 2) tienen poderes causales, sobre otras actitudes o sobre la conducta, y 3) las generalizaciones implícitas de la psicología de creencia/deseo del sentido común son ampliamente verdaderas de ellas. Según lo primero, actitudes proposicionales como los deseos, creencias o corazonadas son verdaderos o falsos, y ello en relación con el mundo. De acuerdo con lo segundo, podemos hablar de causación mental, en el sentido de que la conducta es causada por estados mentales y también, lo cual es más interesante para Fodor, en el sentido de que unos estados mentales son causados por otros estados mentales; asimismo

hay un tercer tipo de causación mental, la de los estados mentales por sucesos del ambiente afectante, esto es, mediante estimulación próxima. Justamente cualquier psicología que es realista acerca de lo mental reconoce *ipso facto* los poderes causales de los estados mentales (frente al conductismo, que rechaza tales poderes causales). Finalmente, según el tercer punto, una buena parte de las generalizaciones de la psicología del sentido común son verdaderas en relación con las actitudes proposicionales, entendidas tal como se establece en los dos puntos anteriores.

En cuanto a la relación entre psicología popular y psicología científica (que para Fodor es obviamente la psicología cognitiva) este autor defiende que es posible una psicología científica que reivindica la explicación de creencia/deseo del sentido común. Tal psicología científica asume la teoría representacional de la mente (una de las tesis básicas de Fodor) o TRM. Esta teoría postula la existencia de un lenguaje del pensamiento, esto es, un conjunto infinito de representaciones mentales que funcionan como los objetos inmediatos de las actitudes proposicionales y como los dominios de los procesos mentales. De un modo más preciso TRM es la conjunción de dos tesis.

La primera tesis especifica la naturaleza de las actitudes proposicionales y dice: para cualquier organismo O y cualquier actitud A hacia la proposición P, existe una relación R (funcional/computacional) y una representación mental RM tal que, RM quiere decir P y O tiene A si y sólo si O tiene la relación R con RM. Es decir, podemos aclarar, existe una representación mental del objeto de una actitud proposicional y el organismo tiene tal actitud proposicional si y sólo si el organismo mantiene una relación funcional/computacional con aquella representación mental.(Como veremos en otro capítulo, la psicología cognitiva postula que los procesos mentales están gobernados por procedimientos computacionales). En suma, tener una actitud proposicional es tener una relación funcional/computacional con una representación mental.

A su vez, la segunda tesis establece la naturaleza de los

procesos mentales y dice: los procesos mentales son secuencias causales de muestras (*tokenings*) de representaciones mentales. Por ejemplo, una cadena de pensamientos, dice Fodor, es una secuencia causal de muestras de representaciones mentales que expresan las proposiciones que son los objetos de los pensamientos.

En resumen, la psicología popular debe ser reivindicada e incorporada a una psicología científica de corte mentalista y método computacional, tal cual ocurre con la psicología cognitiva.

Simplemente quisiera apuntar dos observaciones críticas a la TRM de Fodor. Por un lado, deseos y creencias son ante todo representaciones mentales que no necesitan para su existencia de una proposición explícita y que se cualifican como deseos o creencias por su específico papel como representaciones. Y por otro lado, y en conexión con lo anterior, no existe un nivel de actitudes y otro de proposiciones mentales sino que unas y otras constituyen un único nivel de representaciones mentales.

4. Reservas y ataques a la psicología popular

No todos los autores defienden la psicología popular, tal como hace Fodor, sino que algunos, como Stich, dudan de su fiabilidad, y otros, como Paul Churchland, la rechazan completamente.

Stephen Stich, tanto en su artículo "Autonomous Psychology and the Belief-Desire Thesis" (1978) como en su libro *From Folk Psychology to Cognitive Science* (1983), considera que la psicología popular no debe desempeñar un papel importante en la psicología científica. En efecto, en una psicología científica que acogiera la psicología popular debería estar presente la tesis de creencia-deseo y, al mismo tiempo, lo que denomina Stich "principio de autonomía psicológica", pero ambos elementos entran frecuentemente en conflicto según este autor.

La tesis de creencia-deseo no es sino la asunción popular de que la acción humana debe explicarse, al menos en parte,

en términos de creencias y deseos. A su vez, debo hacer observar que el principio de autonomía psicológica de Stich no tiene el significado fuerte que podría esperarse según la propia denominación, ya que tiene un sentido reductivo de los procesos mentales a los procesos físicos internos del organismo. Tal principio establece que los procesos que deben interesar al psicólogo son aquéllos que sobrevienen (*supervene*) al estado físico interno del sujeto. Con ello se sostiene que los procesos mentales están completamente determinados por los estados físicos correspondientes. Una réplica exacta (átomo por átomo y molécula por molécula) de mí sería psicológicamente idéntica a mí, cualquier propiedad psicológica instanciada por mí sería también instanciada por mi réplica.

Ahora bien, según Stich, la tesis de creencia-deseo de la psicología popular y este principio de autonomía psicológica mantienen tal tensión que las nociones psicológicas populares no deben desempeñar papel alguno en las teorías de la ciencia cognitiva actual. Me limitaré a señalar uno de los ejemplos de esta tensión (mejorando la exposición) aducidos por Stich. Supongamos que en el pasado he probado una botella de *Chateau d'Yquem 1962* y que ahora se construye una réplica de mí. Si se nos pregunta a mí y a mi réplica si hemos probado tal vino diremos que sí, entendiéndose (según la tesis de creencia-deseo) que nuestras respectivas creencias son causas de nuestras respectivas proferencias afirmativas; sin embargo, mientras mi creencia es verdadera la creencia de mi réplica (que acaba de ser creada y no ha probado vino alguno) es falsa. Pero este diferente valor de verdad de tales creencias contraría el principio de autonomía psicológica, según el cual mi réplica y yo debemos instanciar las mismas propiedades psicológicas.

Simplemente apuntaré una observación crítica a este argumento. No es extraño que el principio de autonomía psicológica de Stich entre en conflicto con la tesis de creencia-deseo, puesto que en la psicología popular se asume un dualismo entre procesos mentales y procesos físicos (como veremos más adelante) que es excluido por el principio de autonomía de Stich.

En el caso de Paul Churchland nos encontramos con una

posición más radical contra la psicología popular. En varios de sus escritos, y en particular en su artículo "Eliminative Materialism and the Propositional Attitudes" (1981), Churchland considera que la psicología popular es realmente una teoría, pero una teoría falsa. A su vez, el materialismo eliminativo que propone constituye un programa de investigación conducente a sustituir la psicología popular, tanto en sí misma como en cuanto presente en la ciencia cognitiva, por una neurociencia debidamente completa.

Para Churchland, la estructura psicológica de nuestro sentido común es una concepción falsa y radicalmente engañosa de las causas de la conducta humana y de la naturaleza de la actividad cognitiva. En primer lugar, la psicología popular fracasa ampliamente en la explicación y predicción de los procesos mentales (no sabemos lo que es el sueño, no comprendemos cómo se desarrolla el aprendizaje, no tenemos la menor idea de cómo funciona la memoria, no conocemos lo que es la enfermedad mental), de modo que las cosas más centrales acerca de nosotros permanecen casi enteramente misteriosas dentro de la psicología popular. En segundo lugar, si la psicología popular ha sobrevivido hasta ahora no es a causa de su carácter correcto sino porque los fenómenos mentales son tan difíciles de explicar que cualquier recurso, aunque sea débil, consigue mantenerse. Y en tercer lugar, una neurociencia que no refleje la psicología popular tendrá mucho más éxito explicativo que una neurociencia que refleje la estructura psicológica de nuestro sentido común. En consecuencia, eliminemos la psicología popular y apostemos por la neurociencia.

Creo que Churchland argumenta dialécticamente, en el sentido de que inventa una psicología popular muy ambiciosa para a continuación imputarle fracasos en tareas que la psicología popular real no emprende. En concreto, tal como señalé al principio, las generalizaciones de la psicología popular no son, ni pretenden ser, leyes estrictas, sino mecanismos útiles de explicación y predicción que funcionan en casos sencillos, con lo cual, frente a Churchland, la psicología popular no es una auténtica teoría.

5. Conclusiones

Si nos preguntamos si debemos seguir defendiendo la psicología popular, a pesar de las citadas reservas y ataques, creo que debemos responder afirmativamente, aunque siendo conscientes de las limitaciones y del valor precientífico de la psicología popular.

En primer lugar, las generalizaciones de la psicología popular constituyen un depósito de conocimientos compartido por todos los seres humanos, con una alta capacidad explicativa y predictiva en los casos cotidianos de nuestras vidas (aunque no en los casos extraordinarios o anormales). En segundo lugar, el marco conceptual de la psicología popular (esto es, las nociones de creencia, deseo, sentimiento e intención) es indispensable en buena parte del lenguaje de la psicología científica. Y en tercer lugar, la actitud mentalista propia de la psicología popular tiene plena vigencia con el desarrollo de la psicología cognitiva actual. En suma, podemos hablar de una prolongación natural desde la psicología popular a la psicología científica, paralela a la prolongación desde la lógica natural a la lógica científica. La psicología científica, una vez superado el conductismo, contendrá conceptos y generalizaciones de la psicología popular, aunque dejará de lado algunos elementos suyos, transformará otros e incorporará, por supuesto, nuevas construcciones más sofisticadas para ajustar la explicación psicológica a los casos menos triviales o anormales.

De todas formas, lo que la psicología científica no deberá acoger de la psicología popular es el dualismo ingenuo o acrítico. En efecto, en la psicología del sentido común está incorporado un *cartesianismo popular*, es decir, la tesis de la distinción entre cuerpo y mente junto con la aceptación acrítica de una mente inmaterial. La psicología científica, o bien su filosofía, debe plantearse en términos críticos, de discusión libre y racional, el problema de la relación entre procesos mentales y procesos físicos (problema mente-cuerpo). Creo que la radical hostilidad de Paul Churchland a la psicología popular tiene como fundamento último la postura

materialista de este autor, totalmente incompatible con el dualismo ingenuo presente en la psicología del sentido común.

Por otra parte, el materialismo eliminativo de Churchland, tal como señalamos antes, propone sustituir la psicología en general por la neurociencia, suprimiendo así cualquier residuo de psicología popular que pudiera subsistir en la psicología científica. En su libro *Matter and Consciousness* (1988, edición revisada), Paul Churchland defiende la idea de que cuando la neurociencia haya madurado y se haya establecido la superioridad de su nueva estructura entonces seremos capaces de comenzar a concebir de otra manera nuestros estados y actividades internas. Entonces nuestras explicaciones de la conducta recurrirán a cosas tales como nuestros estados neurofarmacológicos, la actividad neuronal en áreas anatómicas especializadas y cualesquiera otros estados que parezcan relevantes para la nueva teoría. Y entonces también nuestra introspección privada será transformada y podrá ser profundamente aumentada gracias a la estructura más precisa y penetrante con la que será tratada. Según Churchland, todo esto supondrá una revolución conceptual de enorme magnitud así como grandes beneficios para la humanidad.

El materialismo eliminativo va unido, tal como se indica en *Matter and Consciousness,* al materialismo metodológico. Esta metodología practica una aproximación de abajo hacia arriba (*bottom-up*) en lugar de la aproximación de arriba hacia abajo (*top-down*). La psicología cognitiva habitual, señala Churchland, sigue una aproximación de arriba hacia abajo porque comienza con nuestra comprensión corriente de lo que hacen las criaturas inteligentes y luego se pregunta qué tipo de operaciones subyacentes podrían producir o explicar tales actividades cognitivas. En cambio, el materialismo metodológico que va unido al materialismo eliminativo sigue una aproximación contraria de abajo hacia arriba, ya que sostiene la idea básica de que las actividades cognitivas son en última instancia actividades del sistema nervioso, de tal manera que comenzaremos por comprender el comportamiento físico, químico, eléctrico y de desarrollo de las neuronas, y en especial de los sistemas de neuronas, y luego podremos

comprender todo lo que hay que saber acerca de la inteligencia natural.

Esta propuesta de eliminar la psicología en favor de la neurociencia es fuertemente seductora, puesto que supone sustituir las categorías mentalistas propias de la psicología, que son difíciles de manejar, por las categorías físicas propias de la neurociencia, que parecen más concretas y manejables. Sin embargo, personalmente entiendo que la explicación psicológica no es reducible a la explicación neurobiológica, ni por tanto eliminable en favor de la neurociencia. Aunque fuese posible *deducir* los fenómenos psicológicos a partir de los fenómenos neurológicos, no será posible *explicar* los fenómenos psicológicos neurológicamente. Esta distinción entre deducción y explicación ha sido formulada por Hilary Putnam, y con su análisis terminaremos este capítulo.

En efecto, en "Reductionism and the Nature of Psychology" (1973), Putnam sostiene que la doctrina de la Unidad de la Ciencia es errónea. Según tal doctrina (que, podemos añadir, constituye una tesis básica del neopositivismo) las leyes de las ciencias de "nivel superior", como la psicología o la sociología, son reducibles a las leyes de las ciencias de nivel inferior, como la biología, la química y en última instancia la física de las partículas elementales. Sin embargo, arguye Putnam, del hecho de que la conducta de un sistema pueda *deducirse* de su descripción como un sistema de partículas elementales no se sigue que pueda *explicarse* desde tal descripción.

Nuestro autor aclara su punto de vista con el siguiente ejemplo. Supongamos que tenemos un sistema constituido por dos objetos macroscópicos, que son una tabla con dos agujeros, uno cuadrado de una pulgada de ancho y otro redondo de una pulgada de diámetro, y una estaca cuadrada algo menor de una pulgada de ancho. El hecho a explicar es que la estaca pasa por el agujero cuadrado pero no pasa por el agujero redondo. Una explicación aceptable es que, siendo la tabla y la estaca dos objetos aproximadamente rígidos, la estaca pasa por el agujero que es bastante amplio y no por el que es demasiado pequeño. Ahora bien, en tal explicación la microestructura de ambos objetos es irrelevante.

Supongamos ahora, sigue Putnam, que podemos describir la estaca y la tabla como nubes de partículas elementales (newtonianas, para mayor simplicidad) de modo que la estaca es la nube A y la tabla es la nube B; a su vez, el agujero redondo es descrito como región 1 y el agujero cuadrado como región 2. Finalmente, supongamos que, gracias a una hazaña heroica de cálculo, conseguimos probar que la nube A pasará por la región 2 pero no por la región 1. Putnam no cree que esto sea una auténtica explicación, ya que la deducción del hecho a partir de las posiciones y velocidades de las partículas elementales y de sus atracciones y repulsiones eléctricas no es una explicación. En efecto, los rasgos relevantes de la situación, esto es, que tabla y estaca son aproximadamente rígidos y que un agujero es bastante grande para la estaca mientras que el otro no lo es, no aparecen en la deducción, sino que desaparecen en medio de una masa de información irrelevante.

En general, señala Putnam, ciertos sistemas pueden tener conductas respecto de las cuales su microestructura es ampliamente irrelevante. Por ejemplo, en sociología, dada la microestructura del cerebro y del sistema nervioso, no es posible determinar que existirán relaciones de producción capitalistas, sino que las mismas criaturas pueden existir en producción de mercancías precapitalista, feudal o socialista. En particular, la psicología está tan subdeterminada por la biología como por la física de partículas elementales.

7

Teoría causal de la mente: materialismo y funcionalismo

1. La teoría causal de la mente

En contraposición al conductismo, que sostiene que los fenómenos psicológicos se reducen a conducta y que por lo tanto no tiene sentido hablar de la mente como algo distinto de la conducta, podemos hablar en términos generales de la teoría causal de la mente como aquella doctrina amplia que defiende que la mente es causa de la conducta y por tanto distinta de ella.

Más concretamente, la teoría causal de la mente, en sus diferentes versiones, sustenta que la conducta es efecto de causas mentales y que tales causas son procesos internos; en suma, que la conducta es producida por causas mentales internas. Según ello, por ejemplo, mi creencia de que vendrá Fulano es un estado interno que causa mis acciones de reservarle hotel, enterarme de que vendrá en avión, ir a esperarle al aeropuerto, etc. Para la teoría causal de la mente los procesos mentales son algo interno y distinto de la conducta, y asimismo factores causales de la conducta. Por ello, como fácilmente podrá apreciarse, la teoría causal de la mente suscribe el mentalismo.

Con todo, debe hacerse una precisión importante. Aunque para esta teoría la conducta es siempre efecto de procesos mentales internos, sin embargo los procesos mentales no necesariamente producen conducta, ya que es posible tener procesos mentales internos que no se traducen en conducta alguna.

Podemos señalar dos variantes principales, opuestas entre sí, dentro de la teoría causal de la mente. Por un lado, la teoría materialista de la mente o teoría de la identidad mente-cerebro, y, por otro lado, el funcionalismo. Dicho rápidamente (de momento), para la primera los procesos mentales internos, que son causa de la conducta, son idénticos a los procesos cerebrales; en cambio, para el funcionalismo no es legítima la identificación entre procesos mentales y procesos cerebrales, sino que los procesos mentales internos, que son causa de conducta, son estados funcionales cuyo órgano (por así decir) no es necesariamente el cerebro. También podemos señalar los autores y obras básicos de una y de otra variante de la teoría causal de la mente. Dentro de la teoría de la identidad mente-cerebro los autores más destacados son los australianos David Armstrong, cuya obra fundamental es *A Materialist Theory of the Mind* (1968), y David Lewis, cuyo trabajo básico es "Psychophysical and Theoretical Identifications" (1972). A su vez, dentro del funcionalismo los autores más destacados son los americanos Hilary Putnam, con su importante trabajo "Psychological Predicates" (1967), y Jerry Fodor, con su obra *Psychological Explanation* (1968). (Respecto de Putnam y de Fodor debe precisarse que estos autores han modificado sus ideas iniciales; Putnam sigue aceptando el funcionalismo aunque insiste en la determinación de la psicología por la sociología en lo que llama sociofuncionalismo; Fodor ha evolucionado hacia posturas cognitivistas, tal como ya hemos visto).

2. La teoría de la identidad mente-cerebro: Armstrong

La postura materialista en psicología de David Armstrong aparece ya expuesta y justificada en su artículo "The Nature of Mind" (1966), publicado dos años antes de su obra fundamental *A Materialist Theory of the Mind*, razón por la cual analizaremos los argumentos contenidos en tal artículo.

Armstrong comienza señalando que, puesto que los hombres tienen mente (es decir, perciben, tienen sensaciones, emociones, creencias, pensamientos, propósitos y deseos), la

pregunta por la naturaleza de la mente debe ir unida a la pregunta por la naturaleza del hombre. Y respecto de esta cuestión sobre la naturaleza del hombre debemos atender a las respuestas de la ciencia actual. A pesar de los desacuerdos y divergencias entre los científicos, puede considerarse, según nuestro autor, que la doctrina científica establecida es el punto de vista de que podemos dar una explicación completa del hombre en términos puramente físico-químicos. Tal cosa promete en efecto la biología molecular que pretende desvelar los mecanismos físicos y químicos que están en la base de la vida. En consecuencia, debemos intentar desarrollar, concluye Armstrong, una explicación de la naturaleza de la mente que sea compatible con la tesis de que el hombre no es sino un mecanismo físico-químico. Así pues, nuestro autor propone explícita y abiertamente una explicación materialista o fisicalista de la mente.

Ahora bien, cabe preguntarse por qué conceder mayor autoridad a la ciencia que a la religión, a la filosofía, a la ética, o a la literatura y el arte (si éstas, podemos añadir, difieren de la ciencia). En suma, cabe pedir una justificación del cientifismo (que podemos definir como la postura general que sostiene que en cualquier cuestión la última palabra la tiene la ciencia). Armstrong, de modo coherente y honrado, se plantea en efecto esta pregunta y contesta que la ciencia tiene mayor autoridad que las otras instancias u opciones señaladas porque sólo la ciencia, después de investigación prolongada y extendida durante siglos, ha alcanzado un consenso intelectual, especialmente a partir del siglo XVII, y porque además dispone de un método para decidir las cuestiones disputadas, mientras que las otras opciones no disfrutan ni de tal consenso intelectual ni de tal método.

Sin embargo, y como observación crítica, podemos apuntar que el cientifismo de Armstrong adolece de ingenuidad y de una valoración excesivamente optimista acerca de la naturaleza de la ciencia. En efecto, el consenso intelectual no es algo firme y definitivo en la ciencia, y de hecho ésta no sólo es falible sino que cambia constantemente, y además, como ha puesto de relieve reiteradamente Paul Feyerabend, la ciencia

no dispone ni de un único método ni de un método que asegure la verdad de los resultados. En todo caso, podemos conceder a Armstrong, si ésta fuese su tesis, que la ciencia desde un punto de vista racional es un conocimiento más firme que la religión, la filosofía o el arte, pero sin olvidar la posibilidad de la existencia de límites para la razón.

Asimismo nos interesa reseñar la crítica del conductismo que plantea nuestro autor en el artículo que estamos comentando. Mientras la filosofía tradicional, en especial Descartes, había concebido la mente como un campo interno misterioso que yace detrás de la conducta física o externa de nuestros cuerpos, en cambio el conductismo sostuvo que la mente es conducta o bien disposición a la conducta. Para Armstrong, el conductismo es erróneo en ambas versiones. Si la mente es conducta, entonces quedan sin explicar todos los procesos mentales que no van acompañados de conducta alguna. Y si la mente es disposición a la conducta, tal como insistió Gilbert Ryle, entonces tal disposición debe ser algo real, frente a la tesis de Ryle de que una disposición no es un estado o cambio particular; cuando pienso pero mis pensamientos no producen acción alguna, hay algo que sucede realmente en mí y que constituye mi pensamiento, no tratándose de que hablaría o actuaría si se cumpliesen ciertas condiciones.

Aunque el conductismo, insiste Armstrong, es compatible con el materialismo, es una explicación insatisfactoria de los procesos mentales. El habla y la acción son expresión del pensamiento, no idénticos con él. En general, los procesos mentales no pueden identificarse con la conducta, sino que los procesos mentales son la causa interna de la conducta. Queda clara pues la postura mentalista de nuestro autor, dentro de una teoría causal de la mente. Para Armstrong esta concepción de la mente como causa interna de la conducta no convierte al materialismo o fisicalismo en inevitable pero lo hace posible, no implica un punto de vista puramente fisicalista acerca del hombre pero es compatible con tal punto de vista. De nuevo advertimos la honradez y coherencia de nuestro autor, puesto que en cierta medida admite la posibilidad de una causa interna no-física de la conducta.

Ahora bien, prosigue Armstrong, si nuestra noción de la mente y de los estados mentales no es sino la de una causa, dentro de la persona, de ciertos tipos de conducta, entonces la naturaleza intrínseca de esa causa se convierte en una cuestión científica, y no en una cuestión de análisis lógico. Y es ahora cuando se impone el materialismo. Pues no es en la tesis de la substancia espiritual de Descartes, sino en el veredicto de la ciencia actual, donde encontramos la respuesta adecuada. Tal respuesta es que la única causa de la conducta en el hombre y en los animales superiores es el funcionamiento físico-químico del sistema nervioso central. En suma, podemos identificar los estados mentales con estados puramente físicos del sistema nervioso central. Esta es la teoría de la identidad mente-cerebro.

A su vez, en su libro *A Materialist Theory of the Mind*, Armstrong reitera y desarrolla las tesis anteriores. En esta obra distingue de modo claro y explícito las dos etapas que llevan a la teoría materialista de la mente: en una primera etapa se cumple el análisis causal de los conceptos mentales, es decir, la tesis de que los estados mentales son causa interna de la conducta, mientras que en la segunda etapa se identifican tales estados internos con los estados físico-químicos del cerebro, proporcionando así el materialismo de estado central (reducción de la mente a los estados físicos del sistema nervioso central). Según ello, el libro se articula en tres partes. En la primera parte, se estudian las teorías de la mente, criticando el conductismo así como las doctrinas dualistas de la mente (que defienden algún tipo de distinción entre mente y cerebro) y señalando los aspectos y ventajas de la teoría materialista de la mente. En la segunda parte, que es la más extensa, se desarrolla el análisis causal de los conceptos mentales, considerando los diferentes tipos de procesos mentales. Finalmente, en la tercera parte, se insiste brevemente en la identificación de lo mental con lo físico, de los procesos mentales con los procesos cerebrales.

3. La teoría de la identidad mente-cerebro: Lewis

El otro gran autor, además de Armstrong, que defiende la teoría de la identidad entre procesos mentales y procesos cerebrales es David Lewis, en especial en sus artículos "An Argument for the Identity Theory" (1966) y "Psychophysical and Theoretical Identifications" (1972). Realizaré mi exposición y comentarios ciñéndome al segundo de tales trabajos.

Lewis comienza poniendo de relieve que los teóricos de la identidad psicofísica sostienen a menudo que las identificaciones que anticipan entre estados mentales y estados neuronales (del sistema nervioso central) son en lo esencial como algunas identificaciones teoréticas que admitimos sin controversia, por ejemplo la identificación entre el agua y la fórmula H_2O o bien entre la luz y la radiación electromagnética. Asimismo estas identificaciones teoréticas se conciben usualmente como avances teoréticos que hacen posible simplificar la ciencia total *postulando* leyes-puente que identifican algunas de las entidades discutidas en una teoría con entidades discutidas en otra teoría.

Sin embargo, Lewis considera que la identificación psicofísica que debe defenderse no corresponde a esta descripción de la identificación teorética, ya que una teoría fisiológica adecuada puede *implicar* identidades psicofísicas, y no meramente *postularlas*. En "An Argument for the Identity Theory", en efecto, Lewis concluye la identificación entre estados mentales y estados neuronales de acuerdo con la siguiente implicación. Como primera premisa tenemos la identificación entre estados mentales y los ocupantes de papel causal (según la definición de estado mental, es decir, podemos añadir, según la teoría causal de la mente, esto es, la teoría de que los estados mentales son causa de la conducta). Como segunda premisa tenemos la identificación entre estados neuronales y los ocupantes de papel causal (según la teoría fisiológica, es decir, podemos añadir, ya que la fisiología atribuye papeles causales sobre la conducta a los procesos neuronales). Ahora bien, según la propiedad transitiva de la identidad (esto es, si A=B y C=B, entonces A=C), ambas premisas implican (lógicamente)

que los estados mentales son idénticos a los estados neuronales.

Así pues, si los significados de los nombres de los estados mentales proporcionan realmente la primera premisa (es decir, si los estados mentales se definen como factores causales de la conducta), y si los avances de la fisiología proporcionan la segunda premisa (esto es, si la fisiología muestra que los estados neuronales son factores causales de la conducta), entonces se sigue la conclusión de la identificación psicofísica. Como Lewis acepta ambas premisas, también defiende que los estados mentales son idénticos a los estados neuronales.

Pero tal identificación psicofísica no es una identificación teorética en el sentido usual antes señalado, sino una identificación teorética en un sentido más fuerte, precisamente en el nuevo sentido de que tal identificación psicofísica está implicada por las teorías que la hacen posible, y no simplemente postulada por ellas.

4. Evaluación del materialismo

El materialismo o teoría de la identidad mente-cerebro tiene un indudable atractivo doble, ya que tiene el aspecto de una teoría con aureola de progresismo y porque además promete o cumple una clara simplificación dentro del cuerpo de la ciencia total.

En cuanto a lo primero, oponerse al materialismo parece cosa de curas, de filósofos poco informados de la ciencia actual, o de científicos trasnochados, mientras que defender la teoría de la identidad mente-cerebro parece cosa de personas ilustradas, de filósofos al día, o de científicos de vanguardia. En suma, el materialismo parece ser lo *progre* mientras que contradecir el materialismo parece ser lo *carca*, y ya se sabe que en general si se quiere figurar y aparecer hay que ser progresista. Evidentemente este atractivo del aspecto progresista del materialismo no es, en absoluto, un argumento, ni filosófico ni científico, razón por la cual, tras haberlo señalado, no me molestaré en dedicarle más tiempo.

El segundo atractivo señalado es ya serio. El materialismo promete (aunque no ha cumplido) eliminar el lenguaje

mentalista en favor del lenguaje neurofisiológico, y esto sí que sería un gran avance y un buen progreso. No importaría que los psicólogos se quedasen sin trabajo y que la psicología como ciencia desapareciese del reino o república de los saberes, puesto que habríamos reducido la psicología a neurofisiología (al reducir los procesos mentales a procesos neuronales) y con ello habríamos simplificado el cuerpo de la ciencia total.

Sin embargo, esta reducción no parece posible. Ya señalamos en el capítulo anterior que, aunque fuese posible *deducir* los fenómenos psicológicos a partir de los fenómenos neurológicos, no será posible *explicar* los fenómenos psicológicos neurológicamente. La razón básica es que el lenguaje mentalista (con términos como deseo, creencia, emoción, sentimiento, percepción, sensación, recuerdo, inferencia o introspección) no es reducible al lenguaje neurológico (con términos como neurona, sinapsis, neurotransmisor, fibras aferentes o fibras eferentes). Sin duda alguna, la neurociencia en general puede ayudar a comprender un gran número de fenómenos psicológicos, pero no puede proporcionar explicaciones completas de nuestra vida mental. Y a su vez, la razón profunda de esta situación fáctica es que la experiencia de lo mental es irreducible a cualquier explicación neurológica. Con ello quiero decir que nuestro modo de experimentar y, en consecuencia, de conocer los procesos mentales no puede ser sustituido por una explicación de los procesos neurológicos que puedan subyacer en la producción de los procesos mentales. Para decirlo aun más concisamente, una cosa es mi experiencia de los estados mentales y otra cosa su eventual modo de producción neurológica.

A este respecto, como dice el filósofo neozelandés Keith Campbell, en su libro *Body and Mind* (1984, segunda edición), el dolor puede explicarse como un cambio en el patrón y frecuencia de las descargas de neuronas en el cortex o corteza cerebral, pero yo no soy consciente del dolor como una agitación de disparos de neuronas. Más en general, podemos decir que los procesos mentales son vividos y, por ende, conocidos en términos de fenómenos internos de conocimiento, recuerdo, creencia o volición (esto es, en términos del lenguaje menta-

lista), y no son vividos ni, por tanto, conocidos en términos de procesos físico-químicos del cerebro, aunque el estudio de estos procesos pueda resultar esclarecedor e ilustrativo.

De todas formas, dejaré de momento los argumentos generales para intentar responder a las tesis materialistas de Armstrong y de Lewis.

En el caso de la teoría de la identidad mente-cerebro de David Armstrong, la identidad psicofísica, o identidad entre procesos mentales y procesos cerebrales, es meramente una propuesta alentada por el prestigio de la ciencia. En *A Materialist Theory of the Mind*, nuestro autor señala que la identificación de los estados mentales con los estados físico-químicos del cerebro es una apuesta (*bet*), una apuesta casi tan buena, en el actual estado del conocimiento, como la identificación del gen con la molécula de ADN (ácido desoxirribonucleico). Sin embargo, Armstrong pierde de vista que ambas situaciones no son comparables. El gen era, antes de su identificación con el ADN, la unidad de herencia biológica concebida en términos abstractos e indeterminados, mientras que un proceso mental particular (como un sentimiento de odio) es algo concreto y determinado, empíricamente contrastable. Dicho de otro modo, la identificación de los genes con moléculas de ADN dota de sentido concreto a aquéllos, mientras que la identificación de los procesos mentales con los procesos cerebrales reduce el sentido concreto de los primeros al sentido concreto de los segundos sin que tal reducción se justifique salvo apoyándose en el principio de economía. Con todo, como simple propuesta, la identificación psicofísica es un programa de investigación que deberá justificarse a la larga, pero que no parece aceptable por la razón ya aludida de la escasa viabilidad de eliminar el lenguaje mentalista en favor del lenguaje neurológico.

En el caso de la teoría de la identidad mente-cerebro de David Lewis, la identidad psicofísica aparece ya no como una propuesta sino como una conclusión lógica implicada por premisas supuestamente científicas. Sin embargo, ambas premisas pueden ser puestas en duda. En primer lugar, la identificación entre estados mentales y factores causales de la con-

ducta no es en general verdadera. En efecto, y como se señaló anteriormente, hay estados mentales que no producen efecto alguno en la conducta; por ejemplo, nuestros sentimientos pueden no manifestarse en comportamiento alguno, o nuestras imágenes mentales pueden permanecer perfectamente secretas; en ambos casos incluso no es necesario que estos procesos mentales produzcan efectos en otros procesos mentales. En segundo lugar, la identificación entre estados neuronales y factores causales de la conducta tampoco es en general verdadera. Efectivamente, hay procesos neurológicos que no producen acciones que puedan calificarse de conducta en el sentido que interesa al psicólogo; por ejemplo, en la zona de la corteza cerebral motora responsable de los movimientos de la lengua se registran procesos neurológicos que producen la actividad de la lengua durante la masticación de alimentos, pero esta actividad del sistema digestivo no merece la atención del psicólogo.

Por otra parte, la propia conclusión de Lewis, es decir, la identidad entre procesos mentales y procesos neurológicos se enfrenta desde el desarrollo de las ciencias cognitivas (y en especial de la inteligencia artificial) al argumento que denominaré *argumento antimaterialista de las máquinas* y que discurre del modo siguiente. Tal como veremos en el capítulo próximo, existe evidencia empírica de que algunas máquinas tienen procesos mentales. La ciencia de la inteligencia artificial y su tecnología correspondiente han desarrollado máquinas "inteligentes" que ejecutan procesos que en los seres humanos calificamos de procesos mentales; no sólo hacen complicados cálculos aritméticos y pruebas de teoremas lógicos y matemáticos, sino que también juegan a las damas o al ajedrez e incluso realizan diagnósticos médicos. Y, sin embargo, en estas máquinas no existen procesos neurológicos, por la sencilla razón de que no están constituidas de neuronas o células nerviosas. Siendo así, la identidad entre procesos mentales y procesos neurológicos queda refutada, puesto que se dan procesos mentales enteramente ajenos a los procesos neurológicos. En este punto es interesante poner de relieve que el desarrollo de las máquinas, tarea alentada por el

materialismo mecanicista, se vuelve contra tal materialismo en su tentativa de reducir la mente a los mecanismos físico-químicos del sistema nervioso central.

Volviendo a los argumentos generales, podemos señalar *en los seres humanos* distintos tipos de procesos mentales que no parecen poder admitir una explicación en términos puramente físico-químicos y que, por lo tanto, podrían calificarse de procesos no-físicos o incluso espirituales. En primer lugar, tenemos los fenómenos de autoconciencia. Por "autoconciencia" podemos entender dos cosas bien diferentes: 1) la conciencia de nuestros propios procesos mentales, y 2) la conciencia de nosotros mismos como *sujetos* de los procesos mentales. En el primer caso, tenemos procesos de conciencia introspectivos, que se reducen en última instancia a memoria concreta. Pero en el segundo caso tenemos autoconciencia en sentido propio, es decir, la conciencia de nuestra identidad personal; en virtud de ella sé que soy *el mismo* ahora, en mi juventud y en mi infancia y que esta permanencia de mí mismo también se proyecta hacia el futuro. Tal autoconciencia, en cuanto nos suministra el conocimiento de nuestra identidad, a lo largo del pasado, presente y futuro, parece escapar a cualquier proceso neurológico o físico en general.

Bernard Baars, en su libro *A Cognitive Theory of Consciousness* (1988), considera el sistema ERTAS (*Extended Reticular-Thalamic Activating System*), constituido por la formación reticular del tronco cerebral y del cerebro medio, la capa exterior del tálamo y el conjunto de neuronas que se proyectan desde el tálamo a la corteza cerebral, como el espacio de trabajo (*workspace)* global neuronal responsable de los procesos conscientes. Esta tesis es sumamente prometedora y atractiva, al señalar una base neuronal de los procesos conscientes que pone de relieve el carácter integrado de los procesos neuronales. Sin embargo, el sistema ERTAS puede dar razón de la conciencia en el primer sentido señalado de autoconciencia, es decir, como conjunto de procesos conscientes, pero no en el segundo sentido, esto es, como conciencia de nuestra identidad personal. El propio Baars, cuando en la obra citada se ocupa de la noción del yo (*self*), lo considera

como un contexto (*context*) englobante de experiencia, pero en la doctrina de nuestro autor un contexto es un sistema que constriñe contenidos conscientes sin ser él mismo consciente; por tanto, el yo en el sentido de Baars no se corresponde con la autoconciencia en sentido propio.

En segundo lugar, y ligada al fenómeno de la autoconciencia, también es un fenómeno típicamente humano la formación de un proyecto vital personal (idea en la que insistió José Ortega y Gasset). Según ello, cada ser humano no está obligado a vivir según los dictados de sus instintos y según el imperio de las condiciones de su entorno, sino que puede formar un proyecto de vida que guía su actividad personal. Tal proyecto vital supone la elección de un sentido o finalidad para nuestras vidas tomadas globalmente y, en cuanto viniendo desde el pasado y atravesando nuestro huidizo presente se prolonga en nuestro futuro, no parece responder a cualquier proceso neurológico o físico en general. Dicho de otro modo, nuestro programa genético o genoma determina nuestra constitución física y condiciona nuestras posibilidades de acción, pero la elección de nuestro proyecto de vida supone una sobredeterminación en nuestra actividad que se añade a nuestra dimensión física.

Finalmente, y asimismo ligada a la formación de un proyecto vital personal, también es un fenómeno típicamente humano la libertad entendida en sentido propio. Las voliciones (deseos o intenciones) son obviamente un tipo importante de procesos mentales. Mi deseo de ganar un buen sueldo o mi intención de escribir un buen libro son procesos mentales, caracterizados por ser disposiciones o esquemas para ejecutar acciones. Pero las voliciones pueden ser simples o deterministas o bien libres e indeterministas. Por ejemplo, mi deseo de comer a las tres de la tarde, habiendo permanecido sin ingerir alimentos desde las ocho de la mañana, o mi intención de vengarme de fulanito, quien me ha ofendido gravemente, son simples voliciones determinadas por estímulos y condiciones a los que van normal y habitualmente asociadas. En cambio, mi deseo de hacer una huelga de hambre o mi decisión de perdonar a fulanito sus graves ofensas son voliciones libres,

en cuanto contrarias a la determinación de los estímulos e incluso a nuestro propio talante. Es aquí, en las voliciones indeterministas, donde radica la libertad. Así pues, la libertad o conjunto de voliciones libres son decisiones o deseos que el ser humano adopta contrariando los esquemas habituales de acción y el determinismo de la información operante; por ejemplo, la decisión del prisionero hambriento y sediento de no traicionar a sus compañeros, a pesar de que se le ofrece comida y bebida a cambio de sus denuncias. Pues bien, tales voliciones libres parecen indicar una causalidad no-física (y quizás espiritual) y por ello no son reducibles a procesos puramente neurológicos.

En resumen, existen en los seres humanos varios tipos de procesos mentales (procesos de autoconciencia, procesos de formación de un proyecto vital personal, voliciones libres o indeterministas) que no parecen reducibles a los procesos neurológicos. Y esto implica que es posible distinguir en el ser humano entre mente y espíritu; es decir, hay en los seres humanos procesos mentales de índole no-física.

Por otra parte, y tal como hemos indicado antes y veremos en detalle en el próximo capítulo, existen en las máquinas "inteligentes" procesos mentales que tampoco son procesos neurológicos, aunque son procesos físicos. En consecuencia, la teoría de la identidad mente-cerebro no es válida sostenida en términos generales.

Podemos defender la tesis (aunque volveremos sobre este punto más adelante) de que los procesos mentales pueden pertenecer a tres categorías distintas: procesos mentales-cerebrales (y por tanto físicos), procesos mentales físicos aunque no cerebrales, y procesos mentales no-físicos (y posiblemente espirituales).

5. La invención del funcionalismo: Putnam

Tal como señalamos al principio de este capítulo, la teoría causal de la mente (doctrina que defiende que la mente es causa de la conducta y por ende distinta de ella) tiene dos variantes principales: la teoría materialista de la mente o

teoría de la identidad mente-cerebro, que ya hemos analizado, y el funcionalismo, que estudiamos a continuación.

Hilary Putnam ha sido el primer autor en exponer y justificar las tesis del funcionalismo, concretamente en su artículo de 1967 titulado "Psychological Predicates", luego reeditado bajo el nuevo título de "The Nature of Mental States". En tal trabajo, Putnam centra su discusión acerca de la naturaleza de los procesos mentales en la cuestión de si el dolor es un estado cerebral, con la intención de generalizar la respuesta obtenida a todos los estados mentales. En principio, señala nuestro autor, son posibles dos respuestas a tal cuestión. Por un lado, cabe sostener que decir que los dolores son estados cerebrales es perfectamente significativo, esto es, no envuelve ninguna "extensión de uso" (del término "dolores"). Pero por otro lado, cabe defender que decir que los dolores son estados cerebrales no es significativo, porque envuelve una "extensión de uso" o un cambio de significado" (del término "dolores"). La posición propia de Putnam no está expresada por ninguna de ambas respuestas, ya que las nociones de cambio de significado y de extensión de uso están tan mal definidas que no es posible sostener ninguna de ambas respuestas. En efecto, según nuestro autor, ni el lingüista ni el hombre de la calle ni el filósofo poseen actualmente una noción de cambio de significado que resulte aplicable a casos de identificación de términos, tal como el caso presente (donde se identifican los términos "dolores" y "estados cerebrales"). En estas circunstancias, la respuesta a la cuestión de si el dolor es un estado cerebral debe permitir enunciados de la forma "el dolor es A", donde "el dolor" y "A" no son sinónimos en sentido alguno (esto es, podemos añadir, no se determinan recurriendo a la noción de significado), sino que tales enunciados resultan o no aceptables recurriendo a fundamentos empíricos o metodológicos. Tal es el camino seguido por Putnam.

Nuestro autor sostiene que el dolor no es un estado cerebral, es decir, no es un estado físico-químico del cerebro (y tampoco un estado del sistema nervioso entero), sino que es otro tipo de estado. La hipótesis propuesta, en lugar de la identidad entre procesos mentales y procesos cerebrales, es

que el dolor, o el estado de tener dolor, es un estado funcional de un organismo tomado en su totalidad. Pero para aclarar tal hipótesis Putnam utiliza la noción de autómata probabilístico. A su vez, la noción de autómata probabilístico supone el concepto de máquina de Turing. Por ello, expondré brevemente el concepto de máquina de Turing y luego nos ocuparemos de la noción de autómata probabilístico de Putnam.

En 1936, el matemático y lógico británico Alan Turing (1912-1954) publicó un artículo con el título "On Computable Numbers, with an Application to the Entscheidungsproblem", en el que describió en términos generales cualquier computador, llamándose desde entonces máquina de Turing a cualquier computador así descrito. Una máquina de Turing es esencialmente un dispositivo de examen de una cinta potencialmente infinita; tal cinta está dividida en cuadros, que pueden tener o no un símbolo determinado escrito, y sólo se examina cada vez un cuadro de la cinta. Por otra parte, las operaciones de la máquina se reducen a las cuatro acciones siguientes: imprimir un símbolo en la cinta (lo cual supone, si ya hay un símbolo escrito, borrar éste), borrar un símbolo, desplazarse un cuadro a la izquierda y, por último, desplazarse un cuadro a la derecha. Pero resulta muy importante entender que el curso de acción de la máquina está determinado por cuatro elementos: 1) el estado interno actual de la máquina, 2) el contenido del cuadro que está examinando, 3) la operación realizada por la máquina, y 4) el nuevo estado interno que adopta tras la operación. Así pues, un computador, en la descripción general de Turing, posee estados internos, que deben entenderse como la suma total de la información almacenada en la máquina en un momento dado. Dicho sea de paso, tal concepción de un computador recuerda la tesis mentalista de los procesos psicológicos, ya que explícitamente se establecen estados internos.

En "Psychological Predicates", Putnam propone a su vez la noción de autómata probabilístico. Por tal entiende una máquina de Turing con la peculiaridad de que las transiciones entre los estados internos de la máquina pueden responder a variadas probabilidades en vez de ser deterministas. Un

autómata probabilístico tiene entradas sensoriales y asimismo salidas motoras. Con ello, la tabla de la máquina (esto es, podemos aclarar, el conjunto de especificaciones de su curso de acción) establece, para cualquier combinación posible de un estado interno y un conjunto completo de entradas sensoriales, una instrucción que determina la probabilidad del siguiente estado así como también las probabilidades de las salidas motoras. Podemos comentar que la idea del autómata probabilístico consiste en "simular" un organismo con conducta flexible; es decir, dadas ciertas entradas sensoriales y dado un estado interno del organismo, las salidas motoras y el nuevo estado interno no están determinados de modo fijo, sino que responden a ciertas probabilidades.

Ahora bien, sigue Putnam, un sistema dado empíricamente puede ser al mismo tiempo la realización física de muchos diferentes autómatas probabilísticos. Es decir, cabe añadir, que podemos para un mismo sistema u organismo diseñar varios diferentes autómatas probabilísticos. Por ello nuestro autor se ve obligado a introducir la noción clave de descripción de un sistema. En efecto, la descripción de un sistema S es cualquier enunciado verdadero que señala que S posee distintos estados s_1, s_2,....., s_n que están relacionados entre sí, y con las entradas sensoriales así como con las salidas motoras, por medio de las probabilidades de transición entre los estados dadas en cierta tabla de máquina (esto es, el conjunto de especificaciones de su curso de acción). Pues bien, tal tabla de máquina, mencionada en la descripción del sistema, será la organización funcional del sistema relativa a tal descripción. A su vez, un estado cualquiera s_i, tal que el sistema S está en ese estado s_i en un tiempo dado, será el estado total de S (en ese tiempo) relativo a esa descripción.

Putnam pone de relieve que conocer el estado total de un sistema relativo a una descripción envuelve conocer mucho acerca de cómo se comportará probablemente el sistema, dadas varias combinaciones de entradas sensoriales, pero no envuelve conocer la realización física de tal estado total; en particular, no envuelve conocer si tal estado es un estado físico-químico del cerebro. Podemos apreciar que nuestro

autor entiende que el comportamiento de un sistema u organismo se define ante todo por su organización funcional, siendo de importancia secundaria la constitución física del sistema. En términos de teoría de la computación, podemos decir, de modo aproximado, que lo importante no es tanto el hardware (o arquitectura del sistema) como el software (o programa que ejecuta el sistema).

Después del análisis de estas nociones técnicas previas, volvemos a la hipótesis de Putnam de que el dolor es un estado funcional de un organismo. Tal hipótesis, siguiendo a nuestro autor, puede desarrollarse en los cuatro puntos siguientes: 1) todos los organismos capaces de sentir dolor son autómatas probabilísticos, 2) cualquier organismo capaz de sentir dolor posee un tipo apropiado de organización funcional, 3) ningún organismo capaz de sentir dolor posee una descomposición en partes que separadamente posean la citada organización funcional (con ello se indica, podemos añadir, que el dolor es un estado funcional del organismo entero), y 4) para cualquier organización funcional, existe un subconjunto de entradas sensoriales tal que un organismo con la citada organización funcional tiene dolor cuando y sólo cuando algunas de sus entradas sensoriales están en tal subconjunto. En resumen, el dolor es un estado funcional y, en general, los procesos mentales son estados funcionales.

Putnam añade que la organización funcional debe incluir algo parecido a una función de preferencia (y con ello, podemos precisar, la posibilidad de seleccionar entradas sensoriales y salidas motoras, ya que el autómata es probabilístico), asimismo debe incluir algo parecido a una lógica inductiva, es decir, el organismo debe ser capaz de aprender de la experiencia, y finalmente el organismo debe disponer de órganos sensoriales.

Aunque tal hipótesis, que identifica procesos mentales y estados funcionales, pueda parecer vaga, es para nuestro autor mucho menos vaga que la hipótesis materialista, que identifica procesos mentales y estados físico-químicos del cerebro. De hecho, la hipótesis funcionalista es para Putnam más plausible, ya que permite más investigación matemática y empírica y porque conduce a cuestiones y predicciones

fructíferas. Por otro lado, la hipótesis funcionalista, según declaración expresa de nuestro autor, no es incompatible con el dualismo, puesto que un sistema, si lo hubiera, compuesto de cuerpo y alma puede ser un autómata probabilístico. Finalmente, y contra la tesis conductista, la conducta del organismo dependerá no sólo de las entradas sensoriales sino también de su estado total, incluyendo valores, creencias, etcétera.

6. Microanálisis y análisis funcional en Fodor

Además de Hilary Putnam, el otro gran autor que ha propuesto el funcionalismo es Jerry Fodor, al cual ya nos referimos al discutir la psicología popular. Las doctrinas de este autor sobre el funcionalismo se encuentran en especial en su libro de 1968 *Psychological Explanation*, en el cual centraremos nuestro estudio.

Para Fodor, el reduccionismo de las teorías psicológicas a las teorías neurológicas es básicamente erróneo, aunque sea positivo relacionar los constructos psicológicos con los constructos neurológicos. El materialismo reduccionista se caracteriza por practicar el microanálisis en cuanto opuesto al análisis funcional. En el análisis reduccionista o microanálisis la pregunta usual es "¿de qué consta X?", buscando una respuesta que consista en la especificación de la microestructura de X. Por ejemplo, son cuestiones de microanálisis las preguntas "¿de qué consta el agua?" o bien "¿de qué consta un rayo de luz?", siendo sus respuestas respectivas en microanálisis "de dos átomos de hidrógeno unidos a uno de oxígeno" y "de una corriente de electrones".

Contrariamente, el análisis funcional pregunta acerca del papel que juega una parte de un mecanismo en las actividades características de tal mecanismo considerado como un todo. Por ejemplo, una cuestión de análisis funcional es la pregunta "¿qué hace el árbol de levas?" y su respuesta es "abre las válvulas, permitiendo la entrada de combustible en el cilindro, para que al hacer explosión se mueva el pistón".

Por tanto, dice Fodor, mientras el progreso del microanálisis depende del desarrollo de poderosos instrumentos de

observación y de métodos muy precisos de disección, en cambio el análisis funcional requiere una estimación de los tipos de actividad característicos de un determinado mecanismo así como de la contribución de cada una de las partes al funcionamiento del todo. Por otro lado, el microanálisis y el análisis funcional constituyen dos formas muy diferentes de establecer relaciones tanto entre las teorías científicas como entre las descripciones del lenguaje ordinario, razón por la cual no deben confundirse. Cuando consideramos un determinado aparato, por ejemplo un árbol de levas, podemos referirnos a su estructura física y, por tanto, lo entendemos como susceptible de una descomposición en partes. Si en vez de ello hablamos de tal aparato como un elevador de válvulas, entonces nos referimos a su función o papel, sin que interese ahora su posible descomposición en partes. En el primer caso hacemos microanálisis mientras que en el segundo caso hacemos análisis funcional.

Fodor admite y defiende establecer relaciones entre los constructos psicológicos y los constructos neurológicos, pero entiende que se podrán evitar muchas confusiones si, al interpretar los enunciados que relacionan los constructos psicológicos y los constructos neurológicos, no los concebimos como expresión de un microanálisis, sino como una forma de atribuir ciertas funciones psicológicas a los correspondientes sistemas neurológicos. Por ejemplo, añade nuestro autor, para algunos filósofos y psicólogos es posible seguir la trayectoria de una entrada o input a través de los sistemas neurológicos aferentes, centrales y eferentes sin que nos encontremos con nada parecido a motivos, impulsos, estrategias y demás invenciones de las teorías psicológicas. Pero esto es tan erróneo como afirmar que en la descripción completa de un motor de combustión interna nunca aparecerá un elevador de válvulas. En ambos casos se confundirían los términos propios del microanálisis con los términos propios del análisis funcional.

Comentando estos puntos de vista, la identificación entre procesos mentales y procesos cerebrales, propia de la teoría reduccionista, sería análoga a la identificación entre elevadores de válvulas (que se caracterizan funcionalmente) y árboles

de levas (que se caracterizan refiriéndonos a su estructura física). Sin embargo, no todo elevador de válvulas es un árbol de levas, ya que existen otros mecanismos que cumplen la función de elevar válvulas; y análogamente, podemos añadir, no todo proceso mental, que se caracteriza por el rol causal o función que desempeña, es un proceso cerebral, puesto que, además del cerebro, existen otros mecanismos que cumplen funciones mentales, como, por ejemplo, un computador al establecer la demostración de un teorema.

Keith Campbell, en el ya citado libro *Body and Mind* (1984), se pregunta si es legítimo negar la existencia de actividad mental en algo hecho de chips de silicio y alambres de cobre (es decir, en un ordenador) si puede ejecutar tan bien como un ser humano tareas de calcular, resolver problemas o diagnosticar. Para este autor, siguiendo los argumentos funcionalistas, lo que importa para la posesión de características mentales no es de qué está hecha una realidad sino lo que hace y cómo lo hace. Si es la función lo que importa, y la estructura física (o incluso espiritual) subyacente que hace posible las funciones puede variar según los casos, entonces seres humanos, perros, computadores y marcianos (aunque hechos de silicio) pueden tener procesos mentales.

7. Funcionalismo y el problema de los qualia

El funcionalismo es una teoría general de los procesos mentales que ha sido ampliamente aceptada entre los filósofos, los psicólogos y los científicos de inteligencia artificial. Para el funcionalismo, los procesos mentales son procesos internos con un rol causal sobre la conducta, que constituyen funciones mediadoras entre entradas sensoriales y salidas motoras, y que se definen independientemente de su soporte físico. Tal concepción de los procesos mentales resulta útil tanto a psicólogos cognitivos como a los científicos de la inteligencia artificial (siendo la psicología de orientación cognitiva y la inteligencia artificial las ciencias cognitivas principales). Pero además esta concepción también resulta útil al filósofo en cuanto supone lo que podemos calificar de

"tiempo muerto" en la discusión sobre si la mente humana es o no es una realidad exclusivamente física.

Sin embargo, el funcionalismo se enfrenta a una seria objeción. En efecto, para la teoría funcionalista un estado psicológico es idéntico a un estado funcional; en consecuencia, el estado psicológico de experimentar la cualidad de azul deberá ser idéntico a cierto estado funcional; pero puede ocurrir que una experiencia de rojo desempeñe el papel o función correspondiente a la experiencia de azul. En tal caso el estado psicológico ya no puede identificarse, como quiere el funcionalismo, con un estado funcional. Tal es el problema de los qualia o caracteres cualitativos de las sensaciones.

El propio Hilary Putnam, en su libro *Reason, Truth and History* (1981), concretamente en el capítulo titulado "Mind and Body", se plantea e intenta resolver esta objeción. Catorce años después de "The Nature of Mental States" (artículo que hemos expuesto y comentado anteriormente), Putnam se inclina a pensar que el funcionalismo es una teoría correcta. Para él, un espíritu incorpóreo podría presentar cierto programa, un cerebro podría presentar cierto programa, una máquina podría presentar cierto programa, pero la organización funcional de los tres (el espíritu incorpóreo, el cerebro y la máquina) podría ser exactamente la misma aun cuando su materia, su substancia, sea totalmente diferente. Y la razón de ello es que un estado psicológico se identifica y define como un estado funcional.

Ahora bien, sigue Putnam, la teoría funcionalista de la mente halla dificultades con el carácter cualitativo de las sensaciones. Cuando se piensa en un estado psicológico como el de estar enfadado su definición o identificación con un estado funcional de todo el sistema parece muy plausible, pero cuando se piensa en un estado psicológico como el de experimentar una tonalidad particular de azul su identificación con un estado funcional no es plausible. El ejemplo, que ya aparece en Locke, del "espectro invertido" desafía la identificación entre estado mental y estado funcional. Imaginemos (y, podemos añadir, los casos reales de daltonismo, en los que no se perciben determinados colores, hacen la situación algo

fácilmente imaginable) que un individuo ve las cosas de tal modo que las cosas azules le parecen rojas y las cosas rojas le parecen azules; sin embargo, imaginemos que tal individuo ha sido adiestrado desde su infancia de tal manera que cuando ve rojo llama "azul" a ese color y cuando ve azul llama "rojo" a tal color; en estas circunstancias, las demás personas no saben de su anormalidad, ya que habla de azul y habla de rojo como nosotros, aunque percibe respectivamente rojo y azul contrariamente a nosotros.

El problema en esta situación, podemos comentar, es que si la sensación de azul, pongamos por caso, se define o identifica con su función de señalar la presencia de cosas azules y, por otro lado, en un sujeto que padece espectro invertido es su sensación de rojo (y ya no de azul) la que desempeña la función de señalar la presencia de cosas azules, entonces ya no podemos caracterizar las sensaciones de colores como estados funcionales, quedando así refutado el funcionalismo. Dicho brevemente, la función no coincide con la cualidad porque cabe la misma función para distintas cualidades.

Putnam resuelve la objeción haciendo, en este caso de los qualia o caracteres cualitativos de las sensaciones, una concesión a la teoría materialista. Para él, si estos casos son realmente posibles entonces el funcionalista debe admitir que el carácter cualitativo de una sensación es precisamente su materialización física (y no una función); con ello el funcionalista afirma que para este tipo especial de propiedades psicológicas, los qualia, la formulación correcta es la de la teoría de la identidad mente-cerebro.

Posteriormente, en su libro *Representation and Reality* (1988), Putnam ha modificado su actitud funcionalista, aunque sin abandonarla enteramente. La novedad fundamental es que aquí nuestro autor piensa que no podemos caracterizar los procesos mentales sin referencia al ambiente y, en concreto, sin referencia al ambiente tanto social como no-humano. La nueva forma que adopta ahora el funcionalismo es el sociofuncionalismo. Pero para ello Putnam renuncia en gran medida al mentalismo, que fue, podemos añadir, una victoria positiva sobre el conductismo.

8

Inteligencia natural e inteligencia artificial

1. La cuestión de Turing

Tal como hemos visto en el capítulo anterior, según el funcionalismo los procesos mentales son definibles independientemente de su soporte físico o arquitectura, es decir, sin considerar lo que los informáticos denominan hardware en un ordenador. Los procesos mentales, según las tesis funcionalistas, son análogos (o incluso de algún modo equivalentes) a lo que los informáticos llaman software en un ordenador, esto es, al conjunto de programas ejecutables por un ordenador. Ahora bien, tal comparación entre procesos mentales y programas de ordenador plantea la cuestión de si es posible que una máquina tenga procesos mentales e incluso de si una máquina puede "pensar".

Lo cierto es que, aún antes del surgimiento del funcionalismo, el matemático y lógico inglés Alan Turing (del cual ya hemos hablado) se planteó la cuestión de si pueden pensar las máquinas, en un célebre trabajo titulado "Computing Machinery and Intelligence", publicado en la revista *Mind* en 1950.

Turing comienza este artículo señalando que la cuestión de si pueden pensar las máquinas exigiría, en principio, definir previamente los términos "máquina" y "pensar", lo cual, podemos añadir, sería tarea inabarcable ya que podría conducir a una discusión filosófica sin fin. A fin de evitar definir esos términos, nuestro autor propone un juego, que denomina juego de imitación, el cual permite formular una

cuestión inequívoca que es equivalente, según Turing, a la cuestión inicial de si pueden pensar las máquinas. Veamos, pues, en qué consiste el juego de imitación y cuál es esa nueva cuestión equivalente a la primera.

El juego de imitación, en su forma inicial, comprende tres jugadores: un hombre, una mujer y un interrogador (que puede ser hombre o mujer). El interrogador no ve a los otros dos jugadores e incluso, para el mejor funcionamiento del juego, recibe sus respuestas escritas a máquina. El objeto del juego consiste en que el interrogador acierte quién es el hombre y quién es la mujer, utilizando para ello las respuestas de los dos jugadores a sus preguntas. Por otro lado, la dificultad del juego estriba en que mientras la mujer ayuda al interrogador en cambio el hombre trata de engañarle. Por ejemplo, podemos decir, si el interrogador pregunta a ambos jugadores si pueden dejarse la barba, la mujer dirá que no puede pero el hombre mentirá y dirá también que no puede dejarse la barba.

Ahora bien, Turing propone introducir en el juego la modificación consistente en sustituir al hombre por una máquina. En esta nueva forma del juego de imitación podemos plantear la cuestión de si el interrogador se equivoca con la misma frecuencia que cuando jugaba el hombre. Tal es la cuestión inequívoca que, según nuestro autor, equivale a la cuestión de si puede pensar una máquina. En efecto, podemos añadir, si la frecuencia de errores del interrogador es sensiblemente la misma cuando se enfrenta a una máquina engañadora y cuando se enfrenta a un hombre engañador, entonces la máquina puede "pensar" igual que el hombre.

La respuesta del propio Turing es que dentro de unos cincuenta años (es decir, para el año 2000, ya que su artículo es de 1950) será posible programar computadores con una alta capacidad de memoria que jueguen tan bien el juego de imitación que un interrogador normal será engañado por la máquina en el treinta por ciento de los casos. En concreto, nuestro autor cree que a finales de este siglo se hablará del pensamiento de las máquinas sin que nadie se oponga a ello.

Pero debemos hacer algunos comentarios para compren-

der mejor el juego de imitación en su versión final (también llamado test de Turing). Aunque Alan Turing pretendía evitar, con su juego de imitación, discutir las nociones de máquina y de pensamiento, lo cierto es que el propio juego asume de modo explícito una noción de máquina y también asume, aunque de modo implícito, una noción de lo que es pensar. En efecto, la máquina que interviene en el juego sustituyendo e imitando al hombre no es una máquina cualquiera sino precisamente un computador digital universal. Como es sabido, un computador digital consta de tres partes centrales: 1) unidad de memoria, que es el depósito de la información, 2) unidad ejecutiva, que es la parte que realiza las diversas operaciones, y 3) unidad de control, que es la parte que controla que las operaciones se ejecuten correctamente y en el orden adecuado. Además de estas partes centrales la máquina debe poseer (y ello evidentemente es esencial para el juego) una unidad de entrada de datos y una unidad de salida de datos. Por otra parte, un computador digital es una máquina de estado discreto, es decir, que pasa de un estado a otro estado por un salto y no de modo continuo. Finalmente, la máquina implicada en el test de Turing está dotada de universalidad, esto es, puede imitar cualquier máquina de estado discreto, con lo cual puede realizar procesos de computación muy distintos (que, en principio, precisarían de máquinas distintas).

Más importante es la noción de pensamiento implícita en el test de Turing. Tal como se desarrolla el juego de imitación, el "pensamiento" de la máquina que sustituye al ser humano supone comportarse como un ser humano en cuanto a dar muestras de las capacidades de entender preguntas y de dar respuestas así como de engañar al interrogador. En suma, el test indica que una máquina (computador digital universal) piensa si su conducta es comparable a la de una persona hasta el punto de que el interrogador no distingue entre máquina y persona.

2. La inteligencia artificial y sus primeros críticos

Precisamente la ciencia de la inteligencia artificial o, de modo más breve y habitual, la inteligencia artificial se caracteriza por el estudio del diseño y construcción de máquinas inteligentes, entendiendo por tales los mecanismos capaces de realizar tareas que hechas por un ser humano serían atribuidas a su inteligencia.

El desarrollo de la inteligencia artificial fue posible a partir de la creación de los primeros computadores electrónicos en los años cuarenta. Las primeras máquinas de este tipo fueron construidas por el ingeniero alemán Konrad Zuse, antes y durante la Segunda Guerra Mundial, pero tras la derrota alemana su trabajo fue ignorado durante mucho tiempo. Por otro lado, también durante la Segunda Guerra Mundial, un grupo de matemáticos británicos, entre los que estaba Turing, construyeron un computador electrónico llamado Colossus en 1943, cuya tarea principal fue la de descifrar los códigos secretos alemanes. Pero lo cierto es que dado el carácter secreto de estos trabajos tampoco el Colossus fue conocido durante bastante tiempo. En cambio, tuvo más fortuna el primer computador electrónico americano, el célebre ENIAC (Electronic Numerical Integrator and Calculator), construido en 1946 en la Escuela Moore de Ingeniería Eléctrica de la Universidad de Pensilvania. Esta máquina fue ampliamente conocida y divulgada, de tal manera que fue considerada habitualmente durante muchos años como el primer computador electrónico del mundo. Tras estos comienzos, las máquinas se multiplicaron aumentando su velocidad de cálculo, su memoria y su versatilidad, es decir, su capacidad para ejecutar tareas muy diversas.

En los años cincuenta algunos científicos americanos creían de modo firme que los rasgos característicos de la inteligencia humana podían ser descritos en forma tan precisa que pudieran ser simulados por un computador. Con esta idea revolucionaria, un grupo de diez matemáticos y lógicos se reunieron, en el verano de 1956, en el Dartmouth College, en Hanover (New Hampshire). Tal reunión, conocida como la

conferencia de Dartmouth, supuso el punto de partida definitivo de la inteligencia artificial. En ella estaban presentes los más destacados científicos de inteligencia artificial del momento. Arthur Samuel, que había diseñado programas para que un ordenador jugase a las damas, era uno de los asistentes. También estaban Allen Newell y Herbert Simon, quienes presentaron un programa para demostrar teoremas lógicos mediante odenador. Otro participante fue Marvin Minsky, uno de los más entusiastas y gran propagandista de la inteligencia artificial. Finalmente, debe destacarse también la presencia de John McCarthy, quien acuñó la expresión "inteligencia artificial" para referirse a los trabajos que todos ellos estaban realizando. McCarthy fundó en 1957 el laboratorio de inteligencia artificial del Instituto Tecnológico de Massachusetts y, más tarde, en 1963 el laboratorio de inteligencia artificial de la Universidad de Stanford; por otra parte, inventó el lenguaje de computación LISP, en el cual se escribieron buena parte de los programas de computación inteligente.

Desde los años cincuenta el campo de investigación en inteligencia artificial ha avanzado notablemente, con importantes éxitos, aunque también mostrando grandes limitaciones. En primer lugar, se han creado programas de juegos para ordenador, en particular el juego de damas, tres en raya y ajedrez. Ya hemos mencionado los programas del juego de damas de Samuel; en cuanto al ajedrez, los campeones internacionales han sido derrotados en ocasiones por computadores con programas adecuados (como el Chess Champion Mark V). En segundo lugar, se han desarrollado programas para demostrar teoremas lógicos y matemáticos. También hemos mencionado el Logic Theorist, de Newell, Shaw y Simon (publicado en 1956) que probaba la mayoría de los teoremas de los *Principia Mathematica* (obra lógica de Whitehead y Russell); asimismo, por ejemplo, en 1976 se demostró mediante computador el famoso teorema de los cuatro colores. (Tal teorema, que durante más de un siglo se intentó demostrar sin éxito, establece que es posible colorear un mapa político con sólo cuatro colores de modo que los países limítrofes tengan distintos colores).

En tercer lugar, también se han diseñado programas de ordenador para tratar la sintaxis y la semántica de los lenguajes naturales. Existen varios programas de análisis sintáctico y asimismo programas de traducción automática; por ejemplo, Anthony Oettinger creó en 1955 un programa para traducir del ruso al inglés. Sin embargo, las limitaciones en el campo del procesamiento del lenguaje natural son grandes, ya que las reglas sintácticas de los idiomas poseen excepciones y los significados de las palabras varían sensiblemente según los contextos. Otra parcela de investigación en inteligencia artificial es la del reconocimiento de imágenes, como rostros, huellas dactilares u objetos determinados. En 1965, Roberts creó un programa para reconocer objetos tridimensionales que era capaz, por ejemplo, de identificar los objetos en una fotografía. La investigación en reconocimiento de imágenes es importante para la robótica, ya que permite dotar a los robots de la capacidad de ver su situación y los objetos que manejan. Pero también aquí existen limitaciones, puesto que los programas permiten al computador reconocer ciertos patrones de objetos bien definidos pero no otros patrones ajenos al programa ni tampoco objetos de escasa definición.

Finalmente, en esta enumeración no exhaustiva de las parcelas de investigación en inteligencia artificial, se han diseñado con notable éxito los denominados "sistemas expertos", es decir, programas de ordenador que permiten a la máquina dar para problemas específicos soluciones sensiblemente iguales a las que daría el experto humano en tales problemas, pudiendo además la máquina justificar las soluciones ofrecidas. Este es sin duda el campo de mayor éxito de la inteligencia artificial ya que iguala el comportamiento de una máquina a la conducta de un experto humano. En 1964, Joshua Lederberg creó el sistema experto DENDRAL para realizar análisis químicos, sistema que fue perfeccionado en la Universidad de Stanford. Asimismo, en 1968, se diseñó el sistema MACSYMA, que llegó a emular a los expertos matemáticos al poder realizar más de seiscientos tipos diferentes de operaciones matemáticas, incluyendo diferenciación e in-

tegración. Otros sistemas expertos importantes son INTERNIST, que es utilizado para el diagnóstico en medicina interna, MYCIN, que es empleado para el diagnóstico y tratamiento de las enfermedades infecciosas, o el sistema PROSPECTOR, que es usado para la exploración geológica.

Todas estas conquistas de la inteligencia artificial llevaron a muchos de sus autores a creer firmemente que las máquinas podían simular de manera adecuada la inteligencia humana o inteligencia natural, y los argumentos eran la propia evidencia empírica de la existencia de diferentes máquinas inteligentes. Algunos autores incluso fueron más lejos y sostuvieron que estas máquinas no sólo simulaban la inteligencia natural sino que eran realmente inteligentes, con lo que la inteligencia dejaba de ser un atributo reservado al ser humano. Bertram Raphael, en *The Thinking Computer* (1976), llega a afirmar que la evidente inteligencia de los computadores está limitada principalmente por la incapacidad de los seres humanos de programarlos para ser más inteligentes, añadiendo que esta incapacidad se está superando. A su vez, Marvin Minsky ha afirmado que próximas generaciones de computadores serán tan inteligentes que los seres humanos deberíamos estar contentos si estos computadores estuvieran dispuestos a mantenernos en torno a la casa como animalitos domésticos.

Sin embargo, esta creencia de que las máquinas simulan de modo adecuado la inteligencia humana o inteligencia natural así como la tesis más fuerte de que las máquinas realmente piensan han encontrado fuertes críticas.

El mayor y más perseverante crítico de la inteligencia artificial es el filósofo americano Hubert Dreyfus. En 1972 publicó su libro *What Computers Can't Do*, obra que tuvo una edición revisada en 1979. En 1986 volvió a la carga, al publicar con su hermano Stuart Dreyfus el libro *Mind over Machine*. Recientemente ha vuelto a insistir en sus críticas en una nueva edición de su antigua obra, titulada *What Computers Still Can't Do* (1992).

Hubert Dreyfus rechaza no sólo que los computadores realmente piensen sino también que puedan simular de

manera adecuada la inteligencia humana. Sus críticas a la inteligencia artificial se desarrollan poniendo de relieve la falsedad, según él, de las asunciones subyacentes a este nuevo campo de investigación, y asimismo destacando los caracteres específicos del ser humano que no pueden ser imitados por las máquinas.

En cuanto a las falsas asunciones de la inteligencia artificial, Dreyfus denuncia la asunción biológica, la asunción psicológica, la asunción epistemológica y la asunción ontológica que subyacen en la investigación de la inteligencia artificial. La asunción biológica consiste en creer que el cerebro procesa la información en operaciones discretas mediante algún equivalente biológico de los interruptores de encendido o apagado (on/off). La asunción psicológica sostiene que la mente puede considerarse como un mecanismo que opera sobre unidades de información según reglas formales. La asunción epistemológica defiende que todo conocimiento puede ser formalizado, es decir, que todo lo que puede ser comprendido puede expresarse en términos de relaciones lógicas. Finalmente, la asunción ontológica presupone que toda la información relevante acerca del mundo tiene que ser analizable como un conjunto de elementos determinados independientes de la situación.

Para Hubert Dreyfus, estas cuatro asunciones son consideradas por los científicos de inteligencia artificial como algo autoevidente, como auténticos axiomas, mientras que en realidad sólo son meras hipótesis y además cuestionables. La asunción biológica de que el cerebro tiene que funcionar como un computador digital no se ajusta a la evidencia, y las otras tres asunciones conducen a dificultades conceptuales. Sobre la asunción psicológica, nuestro autor señala que hay creciente evidencia de que el procesamiento de información humano y el mecánico proceden de maneras enteramente distintas. Y en cuanto a las asunciones epistemológica y ontológica, que conjuntamente pretenden que toda la conducta humana tiene que ser analizable en términos de reglas (lógicas) que relacionan hechos atómicos, para Dreyfus la formalización de contextos restringidos en inteligencia artificial es una "solución" *ad hoc* que no resuelve el problema de cómo formalizar la

totalidad del conocimiento humano presupuesto en la conducta inteligente.

En cuanto a los caracteres específicos de los seres humanos que no pueden ser simulados por las máquinas, nuestro autor destaca tres aspectos que han sido, según él, descuidados por los científicos de la inteligencia artificial a pesar de que parecen subyacer a toda conducta inteligente. Tales aspectos son el papel del cuerpo en la organización y unificación de nuestra experiencia de los objetos, el papel de la situación para proporcionar una trastienda (*background*) merced a la cual la conducta puede ser ordenada sin ser regulada, y finalmente el papel de los propósitos y necesidades humanas para organizar la situación de modo que los objetos sean reconocidos como relevantes y accesibles.

Dreyfus recurre a filósofos como Martin Heidegger (1889-1976), Ludwig Wittgenstein y Maurice Merleau-Ponty (1908-1961) para apoyar su visión del hombre como una realidad inserta en una situación, determinada por el cuerpo y sujeta a propósitos y necesidades. Desde esta concepción del hombre, la diferencia entre inteligencia mecánica e inteligencia humana aparece como fundamental. La inteligencia artificial abstrae los hechos con los que trabaja de la situación en la que están organizados e intenta usar los resultados para simular conducta inteligente. Pero estos hechos tomados fuera de contexto (fuera de la situación, podemos decir) constituyen una masa dificilmente manejable de datos neutrales. Además los objetivos y necesidades humanas son indeterminados con lo que no pueden ser simulados en una máquina digital cuyo único modo de existencia es una serie de estados determinados. Por ello la inteligencia artificial, en cuanto pretende simular la inteligencia natural o humana, está condenada al fracaso.

Por otra parte, en *Mind over Machine* (1986), Hubert Dreyfus y Stuart Dreyfus (su hermano) evalúan los resultados de la inteligencia artificial considerando que esta investigación ha cumplido pocas de sus promesas y ha fracasado en proporcionar evidencias de que algún día cumplirá sus proyectos. El libro atiende de modo especial al análisis de los

sistemas expertos en inteligencia artificial, gracias a los cuales un computador dotado del programa adecuado puede imitar a un experto humano (químico, matemático, médico o geólogo), para concluir que los computadores no pueden igualar la intuición y pericia humanas. Para los hermanos Dreyfus, los computadores pueden lograr un saber-que (*know that*) pero no alcanzan el saber-cómo (*know how*), esto es, pueden argumentar sobre hechos determinados siguiendo reglas fijas, pero no alcanzan un modo holístico (globalizador, podemos decir) e intuitivo de resolución de problemas. En suma, sería una auténtica estupidez confundir el sentido común, sabiduría y juicio maduro del experto (humano) con la inteligencia artificial actual.

Finalmente, para terminar la exposición de las críticas de Hubert Dreyfus a la ciencia de la inteligencia artificial, me referiré brevemente a las nuevas observaciones que ofrece en *What Computers Still Can't Do* (1992), nueva edición del libro que antes fue considerado. Nuestro autor utiliza la distinción, establecida por John Haugeland en su obra *Artificial Intelligence. The Very Idea* (1985), entre GOFAI (Good Old Fashioned Artificial Intelligence) y otras posibles alternativas de investigación en inteligencia artificial. Según Haugeland, la "buena inteligencia artificial a la antigua usanza" o GOFAI se caracteriza por sostener que los procesos que subyacen a la inteligencia son simbólicos; dicho de modo más explícito, para una teoría GOFAI nuestra capacidad para tratar con las cosas inteligentemente se debe a nuestra capacidad para pensar acerca de ellas razonablemente, y además esta capacidad para pensar sobre las cosas razonablemente se reduce a una facultad de manipulación de símbolos "automática" e interna. En suma, podemos decir, una teoría GOFAI sostiene que la inteligencia es esencialmente manipulación de símbolos interna y automática.

Pues bien, el dictamen de Dreyfus en 1992 es que GOFAI, esto es, la inteligencia artificial clásica, constituye un caso paradigmático de programa de investigación degenerante y que otras alternativas dentro de la inteligencia artificial (como los trabajos en redes neuronales y el aprendizaje por

refuerzo), aunque más interesantes y prometedoras que GOFAI, no parece que puedan simular de modo adecuado la inteligencia humana.

Un programa de investigación degenerante, en la terminología del filósofo de la ciencia Imre Lakatos (1922-1974), es una empresa científica que comienza con grandes promesas al tiempo que ofrece una nueva aproximación que conduce a resultados impresionantes en un dominio limitado. A continuación los investigadores intentan aplicar esa nueva aproximación a nuevos dominios y, en la medida en que se producen éxitos, el programa de investigación se expande y atrae seguidores. Sin embargo, surgen fenómenos inesperados e importantes que resisten las nuevas técnicas y, finalmente, el programa se estanca provocando el abandono de los investigadores en cuanto surge alguna aproximación alternativa. Tal es, según Dreyfus, la historia de GOFAI, es decir, de la inteligencia artificial clásica preconizada por Newell, Simon y Minsky, entre otros. Nuestro autor sostiene que este fracaso de GOFAI fue previsto por él desde 1965, al comprender que nuestro sentido de la relevancia es holístico (globalizador) y que requiere compromiso con la actividad en marcha, mientras que las representaciones simbólicas, propias de GOFAI, son atomísticas y totalmente separadas de tal actividad.

La degeneración del programa GOFAI, según Dreyfus, se completó con la llegada triunfante de los revolucionarios de las redes neuronales o conexionistas. Aunque tendremos otras ocasiones para referirnos a la teoría de redes neuronales o conexionismo, diremos de momento que esta nueva aproximación en inteligencia artificial defiende que las unidades de información se representan mediante elementos muy simples de computación que intercambian mensajes simples, y que las computaciones complejas se realizan mediante redes de esos elementos simples conexionados de manera masivamente paralela. Una máquina conexionista imita el cerebro humano o, más exactamente, las redes de neuronas, y además procesa la información en paralelo y no, como en GOFAI, de modo serial o secuencial.

Hubert Dreyfus se muestra muy interesado en señalar la

idea de Paul Smolensky (expuesta en su artículo de 1987 "Connectionist AI, Symbolic AI, and the Brain") de que en los sistemas conexionistas el conocimiento es codificado no en estructuras simbólicas (como en la inteligencia artificial clásica) sino en la configuración de las fuerzas numéricas de las conexiones entre los procesadores. En efecto, tal idea parece encajar con la tesis de Dreyfus de que los computadores en cuanto sistemas simbólicos físicos no pueden simular la inteligencia. Sin embargo, tampoco la aproximación conexionista, según nuestro autor, ofrece éxitos que puedan generalizarse como una simulación adecuada de la inteligencia humana, ya que en última instancia el problema de simular el conocimiento de sentido común humano tampoco se resuelve en esta nueva tendencia de la inteligencia artificial.

Por último, la aproximación más reciente de inteligencia artificial, a saber, el aprendizaje por refuerzo, tal como aparece en el número de mayo de 1992 de la revista *Machine Learning*, tampoco consigue, según Dreyfus, simular adecuadamente la inteligencia humana. Los teóricos del aprendizaje por refuerzo intentan diseñar sistemas que aprendan por sí mismos cómo enfrentarse con el ambiente y cómo modificar sus propias respuestas conforme cambia el ambiente. Sin embargo, según nuestro autor, una máquina de aprendizaje por refuerzo no consigue dominar el problema de todo lo que es relevante en cada situación. En cambio, los seres humanos disponen de necesidades, deseos e incluso emociones que les proporcionan de modo directo el sentido de lo que es apropiado para cada conducta.

Otro de los primeros críticos de la inteligencia artificial es Joseph Weizenbaum, autor del libro *Computer Power and Human Reason* (1976). Mientras las críticas de Hubert Dreyfus fueron, en general, recibidas con desprecio por los investigadores de inteligencia artificial, considerando que Dreyfus era ajeno e ignorante del campo de investigación que criticaba, en cambio esta actitud de desprecio hacia un extraño no fue posible en el caso de Weizenbaum, ya que éste había hecho notables aportaciones a la inteligencia artificial, por lo que sus críticas fueron, en general, recibidas como las de un

traidor. En efecto, Weizenbaum creó un programa de computador llamado ELIZA, descrito en su trabajo "ELIZA. A Computer Program for the Study of Natural Language Communication between Man and Machine" (1965), diseñado para simular una conversación entre un psicoanalista (papel atribuido a la máquina) y un paciente. El programa no pretendía curar al paciente, aunque sí permitirle explayarse en sus problemas, y la máquina daba respuestas a las preguntas que le formulaba el paciente siguiendo el hilo de su conversación o respondiendo con frases muy generales cuando las preguntas pretendían una respuesta técnica en psicoanálisis.

En el citado libro *Computer Power and Human Reason* (1976), nuestro autor sostiene la tesis general de que el hombre no es una máquina y, aunque ciertamente procesa información, no lo hace como un ordenador necesariamente. De donde podemos concluir, siguiendo a Weizenbaum, que ningún ordenador podrá simular de modo adecuado la inteligencia humana o natural. De modo más concreto, nuestro autor pone de relieve que la inteligencia por sí misma es un concepto sin sentido, siendo preciso un punto de referencia, la especificación de un dominio de pensamiento y acción, a fin de que adquiera significado. Por ello al hablar de "inteligencia" hemos de hacer explícito el dominio de pensamiento y acción que hace inteligible el término, con lo que hablar de inteligencia absoluta (y no relativa a una tarea, podemos añadir) lleva al debate estéril acerca de si es posible construir ordenadores más inteligentes que el ser humano. Para responder, cabe comentar, será preciso tener en cuenta tareas concretas. Como dice nuestro autor, los ordenadores derrotan a campeones de ajedrez y resuelven grandes sistemas de ecuaciones, pero son incapaces de cambiar el pañal de un niño.

Weizenbaum señala las limitaciones de los ordenadores frente a las específicas capacidades del saber humano. En primer lugar, el poder de los ordenadores para obtener conocimiento a partir del lenguaje natural que se les administra se limita a una mínima parte de conocimiento importante, siendo en esto un ordenador inferior a un niño de dos años. En

segundo lugar, el ser humano dispone de conocimientos que adquiere en virtud de poseer un cuerpo humano y que ningún organismo, incluido por supuesto un computador, que carezca de tal cuerpo llegará a adquirirlos de la misma forma. En tercer lugar, hay ciertos hechos que el individuo humano logra saber únicamente como consecuencia de haber sido tratado como ser humano por sus semejantes; aún admitiendo que un ordenador desarrolle un sentido de sí mismo y sea capaz de modificarse por sus experiencias, la socialización humana permanecerá como necesariamente distinta de la socialización de la máquina. Por mucha inteligencia que puedan alcanzar los ordenadores, será una inteligencia *extraña* a las cuestiones y problemas genuinamente humanos.

Por último, el contenido de la información de un mensaje no es sólo una función del mensaje, sino que depende fundamentalmente del estado de conocimiento y de las expectativas de quien lo recibe. Así, el mensaje "Llego en el avión de las siete. Con amor, Bill" tiene un contenido informativo distinto para la esposa de Bill, quien sabía que regresaba aunque no en qué avión, del que tiene para la joven que no esperaba a Bill y que además se ve sorprendida por su declaración de amor. El lenguaje humano, señala Weizenbaum, incluye la historia de quienes lo utilizan y, por tanto, la historia de la sociedad y de la humanidad entera. En cambio, el lenguaje de los ordenadores es sólo funcional, identificando hechos y palabras únicamente con las metas inmediatas que han de lograrse o con los objetos que deben ser transformados. Pues bien, según nuestro autor, estos cuatro tipos de limitaciones no sólo son limitaciones técnicas de los ordenadores, sino que además indican la diferencia entre ser un ser humano o ser un ordenador.

3. Test de Turing y habitaciones orientales

El debate acerca de la inteligencia artificial, en cuanto a sus pretensiones de crear programas de computador que igualen o incluso superen la inteligencia humana o natural, ha continuado con nuevos argumentos en favor y en contra.

El filósofo americano John Searle, en su trabajo "Minds,

Brains, and Programs" (1980), presenta un ingenioso argumento, el experimento mental de la habitación china, para mostrar que es posible superar el test de Turing (es decir, comportarse como si se pensase) y sin embargo no pensar realmente. Antes de exponer tal argumento, Searle distingue entre lo que llama inteligencia artificial fuerte y lo que estima inteligencia artificial débil o prudente. La segunda se caracteriza por considerar los computadores como poderosos instrumentos para el estudio de la mente, permitiéndonos formular y comprobar hipótesis de un modo más riguroso y preciso. En cambio, la inteligencia artificial fuerte considera los computadores como más que un instrumento en el estudio de la mente, ya que sostiene que un computador programado de manera apropiada es realmente una mente al poder literalmente pensar y tener otros estados cognitivos. Nuestro autor no pone objeciones a la inteligencia artificial débil, es decir, podemos añadir, al empleo de los computadores como un recurso instrumental y conceptual para la comprensión de la mente, pero ataca de modo decidido la pretensión de la inteligencia artificial fuerte de que los computadores realmente piensan.

Para John Searle, los partidarios de la inteligencia artificial fuerte sostienen que, en un diálogo entre un ser humano y un computador (como, por ejemplo, el diálogo entre un hombre y una máquina con el programa ELIZA de Weizenbaum, citado antes), la máquina no sólo está simulando una capacidad humana sino que también puede decirse que la máquina literalmente *comprende* y que lo que hace la máquina *explica* la capacidad humana de comprensión. Para refutar tal pretensión nuestro autor presenta el siguiente experimento mental (o imaginado), que denomina argumento de la habitación china.

Supongamos que un hombre, que no sabe nada de chino aunque sí domina el inglés, está encerrado en una habitación. Este hombre dispone de instrucciones en inglés para manejar listas de símbolos en chino y correlacionarlas teniendo en cuenta únicamente las formas de los símbolos, de tal manera que cuando recibe una lista de símbolos en chino entrega otra

lista de símbolos en chino que es una respuesta adecuada a la primera. En estas condiciones, para un observador exterior a la habitación, las respuestas del hombre encerrado a las preguntas en chino no se distinguen de las respuestas que daría un hablante chino. Y sin embargo el hombre encerrado produce sus respuestas simplemente manipulando símbolos no interpretados.

Según Searle, un computador se comporta exactamente igual que el hombre encerrado en la habitación china, es decir, da la impresión de pensar y comprender pero en realidad se limita a manipular símbolos formales que no comprende. El hombre encerrado en la habitación china supera el test de Turing, esto es, se comporta como si pensase en chino, y sin embargo es incapaz de pensar en chino. Del mismo modo un computador digital puede superar el test de Turing, dando la impresión de que piensa, y sin embargo realmente no piensa, puesto que se limita a una manipulación de símbolos formales sin significado alguno, porque sus símbolos sólo tienen sintaxis pero no semántica. En suma, podemos decir, superar el test de Turing no equivale ni implica pensar (frente a las pretensiones del propio Turing).

Para Searle, la única máquina que puede pensar es el cerebro, o bien máquinas que tuvieran los mismos poderes causales que el cerebro. Sobre esta idea vuelve en su artículo "Minds and Brains Without Programs" (1987). En este trabajo resume sus puntos de vista en una serie de "axiomas" y conclusiones derivadas, de los que destacamos los más importantes. El primer axioma establece que los cerebros causan las mentes, con lo que para nuestro autor queda excluida de entrada cualquier mente mecánica. El segundo axioma defiende que la sintaxis no es suficiente para la semántica, marcando así una radical distinción entre el nivel formal y el nivel de los significados. El tercer axioma señala que las mentes tienen contenidos, específicamente contenidos intencionales o semánticos. A su vez, como contrapunto al anterior axioma, Searle defiende en el cuarto axioma que los programas (de computador) se definen de manera puramente formal o sintáctica.

De tales axiomas, que nuestro autor considera puntos obvios (y que al final de este apartado criticaré), extrae algunas conclusiones. De ellas, la más relevante es la siguiente: instanciar un programa (es decir, ser un caso de ejecución de un programa) no es nunca por sí mismo suficiente para tener una mente. Tal conclusión, según Searle, refuta la inteligencia artificial fuerte, esto es, la pretensión de que las máquinas tienen mente. Podemos aclarar este dictamen señalando que la habitación china constituye, para Searle, una instanciación de un programa (el de las instrucciones que correlacionan las listas de símbolos en chino) y sin embargo la habitación china no tiene mente; dicho de otro modo, aunque la habitación china realiza un programa de entender y hablar chino, sin embargo la habitación china no comprende el chino.

El experimento mental de Searle ha sido ampliamente comentado y discutido. Nos limitaremos a exponer las críticas de Rapaport y de Paul y Patricia Churchland.

William Rapaport, en su trabajo "Syntactic Semantics: Foundations of Computational Natural-Language Understanding" (1988), corrige el argumento de la habitación china exponiendo una variación que denomina argumento de la habitación coreana (aunque en realidad aquí no hay habitación alguna). Supongamos un catedrático coreano de literatura inglesa en la Universidad de Seúl que no comprende el inglés, ni hablado ni escrito, pero que no obstante es una autoridad mundial en Shakespeare. Tal autoridad la ha alcanzado leyendo a Shakespeare en excelentes traducciones coreanas y habiendo escrito varios artículos en coreano sobre el autor inglés, artículos que a su vez han sido traducidos al inglés por otras personas y publicados en revistas eruditas de fama, donde han sido recibidos con gran éxito.

Podemos decir que la situación del catedrático coreano respecto de la obra de Shakespeare es comparable a la situación del hombre encerrado en la habitación respecto del chino. La cuestión ahora es si el catedrático coreano "comprende" a Shakespeare y la respuesta es obviamente afirmativa, de tal manera que el profesor coreano comprende algo. De modo similar, según Rapaport, el hombre encerrado en la habita-

ción comprende algo. Podemos añadir que, aunque ni el profesor coreano comprende el inglés ni el hombre encerrado en la habitación comprende el chino, ambos comprenden algo, el uno a Shakespeare y el otro cómo relacionar listas de símbolos chinos (aunque Rapaport va más lejos, insistiendo en que el hombre de la habitación comprende chino). En general, un computador con un programa de comprensión del lenguaje natural supera el test de Turing, es decir, se comporta como si pensase, y ello es una señal de inteligencia. Por tanto, podríamos concluir la existencia de inteligencia artificial, aunque personalmente dudo que tal inteligencia iguale, en el caso de la comprensión del lenguaje natural, a la inteligencia humana.

A su vez, Paul Churchland y Patricia Churchland, en su artículo conjunto "Could a Machine Think?", publicado en *Scientific American* (enero de 1990), también critican el argumento de la habitación china. La tesis general de estos autores es que la inteligencia artificial clásica tiene dificultades para crear máquinas que piensen, pero que en cambio la nueva inteligencia artificial inspirada en la estructura cerebral humana puede crear máquinas que piensen. Con ello volvemos al tema, ya aludido antes, de la distinción entre inteligencia artificial clásica (que llamábamos GOFAI, siguiendo a Haugeland) y la inteligencia artificial conexionista. La primera concibe los computadores como máquinas manipuladoras de símbolos de acuerdo con reglas sensibles a estructuras formales y que trabajan de manera serial o secuencial. En cambio, la inteligencia artificial conexionista concibe los computadores como máquinas de procesamiento masivamente paralelo que imitan el funcionamiento de las redes neuronales.

Para Paul y Patricia Churchland, si analizamos la estructura del cerebro encontramos tres rasgos que contrastan con la arquitectura de los computadores electrónicos convencionales. Para empezar, el sistema nervioso es una máquina paralela, es decir, las señales se procesan en millones de sendas diferentes de modo simultáneo. En segundo lugar, la unidad de procesamiento básico, la neurona, es comparativamente simple y su respuesta a las señales que recibe es

análoga, no digital, en la medida en que varía de modo continuo. Y en tercer lugar, las conexiones que van de una población de neuronas a otra se emparejan a menudo con conexiones de retorno, haciendo así del cerebro un sistema dinámico genuino.

Nuestros autores son partidarios del desarrollo de una inteligencia artificial que explote e imite nuestro conocimiento del funcionamiento del cerebro, sin descartar por ello la existencia de máquinas clásicas, que son útiles para ciertas tareas. Una arquitectura paralela, en vez de serial, de los computadores proporciona una superior velocidad en comparación con un computador convencional, ya que las muchas conexiones en distintos niveles ejecutan simultáneamente muchas computaciones en vez de en forma secuencial. Además, el paralelismo masivo hace al sistema tolerante con los fallos y persistente en su función, pues la pérdida de algunas conexiones tiene un efecto despreciable en la transformación general ejecutada por la red total. También un sistema paralelo almacena cantidades amplias de información de una manera distribuida, siendo accesible cada una en milisegundos.

Finalmente los Churchland destacan que una máquina conexionista no es un sistema manipulador de símbolos según reglas formales. Sin embargo, la crítica de Searle contenida en su argumento de la habitación china va dirigida contra las máquinas manipuladoras de símbolos, que se comparan, podemos añadir, con el hombre encerrado en la habitación y que se limita a manipular listas de símbolos chinos. Al margen de que tal sistema manipulador de símbolos pueda generar significados (lo cual es tajantemente rechazado por Searle, mientras que los Churchland admiten tal posibilidad), en todo caso la inteligencia artificial a la usanza conexionista, imitadora del cerebro humano, promete crear máquinas con auténtico pensamiento según la opinión de nuestros autores.

Para terminar este apartado, señalaré por mi parte algunas otras limitaciones en el argumento de la habitación china. En primer lugar, el argumento supone una distinción tajante entre sintaxis y semántica que no es admisible de modo tan radical. En efecto, podemos distinguir entre, por un

lado, la forma o sintaxis de una cadena de enunciados del lenguaje natural y, por otro lado, su semántica o significados concretos; pero las formas o estructuras del lenguaje son tipos de ordenación o disposición de las palabras en virtud de la cual las frases o cadenas de enunciados tienen una u otra manera general de significar. Por tanto, cada forma o estructura del lenguaje tiene una manera de significar, no siendo posible una distinción tajante entre sintaxis y semántica. Por ejemplo, una conjunción de enunciados se distingue de una implicación de enunciados en cuanto en la primera significamos una acumulación de enunciados mientras que en la segunda significamos que dado un enunciado debe producirse el otro; estas estructuras o formas del lenguaje (conjunción e implicación) son distintas por esa manera distinta de significar, independientemente de los significados concretos de sus enunciados componentes. En suma, las formas o sintaxis prescinden de los significados concretos (y en este sentido es posible distinguir entre sintaxis y semántica), pero las formas del lenguaje no prescinden de sus específicas maneras de significar o decir (y en este otro sentido no es posible la distinción tajante entre sintaxis y semántica).

En segundo lugar, el hombre encerrado en la habitación china produce significados, contrariamente a lo que sostiene John Searle, si admitimos la noción de significado del segundo Wittgenstein. En efecto, Wittgenstein, en la segunda fase de su actividad filosófica, defendió la idea de que el significado de las palabras se define mediante su uso en el lenguaje y no mediante su referencia, puesto que una palabra tiene significado aún en el caso de que nada real corresponda a ella. Las palabras por tanto tienen significado para nosotros si comprendemos su uso. Admitiendo esto, el uso que el hombre encerrado en la habitación hace de las listas de símbolos en chino, según las instrucciones que ha recibido en inglés y que comprende, confiere significado a tales listas de símbolos, con lo que el sistema, frente a Searle, tiene contenido semántico.

Finalmente, el "axioma" de John Searle de que sólo los cerebros producen mente constituye una clara petición de principio, ya que contiene en sí mismo el rechazo de una

máquina que piense, conclusión a la que pretende no obstante llegar más tarde. Además tal axioma supone un ingenuo chauvinismo humano y la creencia en la radical y substancial superioridad de la química del carbono, propia de los seres vivos, sobre la química del silicio, propia de los computadores.

4. Conclusiones

El surgimiento de la inteligencia artificial ha despertado entre los seres humanos pasiones encontradas. Por una parte, algunos científicos han visto en las llamadas máquinas inteligentes nada menos que los sustitutos perfeccionados de la raza humana, mientras que otros científicos consideran que las denominadas máquinas inteligentes son un remedo imperfecto del hombre y permanecerán siempre como tales. Un ejemplo de la primera actitud, de fervoroso entusiasmo, es Robert Jastrow quien, en su libro *The Enchanted Loom* (1981) considera a los computadores como los sucesores evolutivos de los seres humanos. Para Jastrow, destacado científico de la NASA, la humanidad está destinada a tener un sucesor aún más inteligente, las poderosas fuerzas evolutivas (más culturales que biológicas) conducirán a una forma de vida inteligente, más exótica y evolucionada a partir del hombre, pero que será hija de su cerebro y no de sus órganos sexuales; tal nueva forma de vida será el computador.

En general, los filósofos han visto con escepticismo el desarrollo de la inteligencia artificial, actitud unida al sentimiento, más profundo, de que las máquinas inteligentes podrían destronar al hombre de su privilegiada situación en el mundo. Desde los comienzos de la actividad intelectual, el hombre se ha definido a sí mismo como un ser inteligente, excluyendo así la inteligencia no sólo de los otros animales sino también de cualquier máquina imaginable. Y sin embargo su esencia y su puesto central en el mundo están amenazados por la posibilidad de máquinas inteligentes. Ello ha provocado reacciones violentas, como en el caso de Hubert Dreyfus.

Dejando de lado el ámbito de las pasiones (fervoroso entusiasmo o violenta hostilidad), lo cierto es que a lo largo de

la breve historia de la inteligencia artificial, apenas cincuenta años, se han producido notables éxitos y también estruendosos fracasos. La demostración automática de teoremas lógicos y matemáticos, los programas de juegos o los sistemas expertos (que simulan a expertos humanos, como los matemáticos, médicos, químicos o geólogos) han proporcionado notables éxitos. Pero los programas de procesamiento del lenguaje natural o los sistemas de reconocimiento de imágenes han fracasado en gran medida. Tras una primera época de éxitos sorprendentes, los fracasos o éxitos limitados fueron surgiendo conforme el programa de investigación en inteligencia artificial se hacía más ambicioso.

Además la inteligencia artificial no obedece a un único paradigma de investigación. Ya hemos visto que, junto a la buena inteligencia artificial a la antigua usanza (GOFAI), se está desarrollando de modo muy prometedor la inteligencia artificial a la usanza conexionista. Justamente el conexionismo está logrando avances en campos difíciles para la inteligencia artificial clásica, como programas de tratamiento del lenguaje natural, sistemas de reconocimiento visual de objetos, o desarrollo de detectores de rasgos especiales. Asimismo los trabajos en máquinas de aprendizaje por refuerzo también resultan prometedores, y nada excluye nuevos paradigmas dentro de la inteligencia artificial.

Sin embargo, cuando nos planteamos la cuestión de la inteligencia de los computadores (y ya no el éxito o el fracaso de la ciencia de la inteligencia artificial), debemos adoptar el punto de vista, ya señalado, de Weizenbaum de que no tiene sentido una caracterización abstracta o absoluta de la inteligencia, sino que es preciso una estipulación concreta y relativa a las tareas que en el ser humano (punto de referencia histórico) consideramos inteligentes. Entendidas así las cosas, resulta evidente que existe inteligencia artificial, ya que existen máquinas inteligentes, es decir, máquinas con programas para ejecutar tareas que en el ser humano atribuimos a su inteligencia; por ejemplo, máquinas para demostrar teoremas difíciles, o para diagnosticar enfermedades infecciosas, o para resolver complicadas ecuaciones matemáticas.

Pero no es menos cierto que las máquinas, aunque en ciertas tareas igualen e incluso superen a la inteligencia humana, no consiguen igualar al ser humano en todas sus tareas inteligentes. El punto decisivo en esta cuestión es la (aparentemente) insuperable dificultad de las máquinas en igualar algo tan sencillo en el hombre como es su sentido común. En virtud del sentido común, el ser humano tiene unas capacidades intuitivas, que no parecen poder representarse en un procedimiento mecánico, de distinguir lo relevante de lo irrelevante en la información, de advertir lo que cambia y lo que permanece en una situación, y de revisar sus conocimientos y conclusiones según se producen nuevos datos. Precisamente buena parte de la investigación lógica actual en conexión con la inteligencia artificial dirige sus esfuerzos hacia los objetivos de formalizar y mecanizar el razonamiento del sentido común, habiendo surgido una veintena de nuevas lógicas, tal como puede verse en el libro de Witold Lukaszewicz de 1990 *Non-Monotonic Reasoning. Formalization of Commonsense Reasoning*.

Otro punto decisivo en las diferencias entre la inteligencia artificial y la inteligencia humana lo constituye la existencia de procesos "espirituales" en los seres humanos, a los que nos hemos referido en el capítulo anterior. Los procesos mentales de autoconciencia (o conciencia de sí mismo como sujeto de los procesos mentales), los procesos de formación de un proyecto vital personal y las voliciones libres o indeterministas parecen ser características exclusivas de los seres humanos, que muy improbablemente llegarán a poseer las máquinas y que de modo muy plausible suponen una índole no-física (espiritual).

A ello podríamos añadir los fenómenos parapsicológicos, que suscitan muchas dudas y frecuentemente son casos de fraude, pero que en numerosas ocasiones han sido constatados empíricamente. Los fenómenos de telepatía (comunicación de pensamientos a distancia), de precognición (conocimiento intuido de sucesos futuros) o de psicocinesis (movimiento de objetos mediante concentración mental) no parece que puedan ser compartidos con las máquinas. De hecho,

cuando Turing, en su ya citado "Computing Machinery and Intelligence", responde a los argumentos en contra de su tesis de que las máquinas pueden llegar a pensar, sólo considera como argumento en contra de mucho peso los casos de fenómenos parapsicológicos.

En resumen, podemos decir que algunas máquinas dotadas de programas adecuados en efecto piensan y están dotadas de mente, aunque sus procesos mentales, superiores en algunas tareas a los procesos humanos correspondientes, no agotan la enorme variedad y capacidades de tareas de los procesos mentales de los seres humanos. El hombre es y seguirá siendo el paradigma de la inteligencia.

9

El problema mente-cerebro: datos e hipótesis

1. Introducción al problema

Repetidamente, a lo largo de este libro, han aparecido referencias al denominado problema mente-cuerpo. Sin embargo, debemos aclarar las distintas formulaciones de este problema y centrarnos en aquella formulación que más interese al planteamiento actual de la filosofía de la mente, para luego señalar las soluciones de mayor interés o vigencia.

Desde la prehistoria humana el hombre se ha entendido a sí mismo como una realidad dual, al menos usando dos argumentos distintos. Por un lado, distinguía entre su vida actual y una vida imaginada después de la muerte en la que podría sobrevivir algo de su naturaleza. Por otro lado, distinguía entre su cuerpo, en medio de otros cuerpos, y su voluntad, que podía sentirse ajena a su propio cuerpo y no satisfecha con él. Al menos este dualismo entre lo que somos y lo que seremos o podríamos ser es algo presente en el ser humano desde que es *homo sapiens*. Las más primitivas religiones y los mitos más antiguos testimonian esta idea del hombre como realidad dual.

Es a esta dualidad apercibida o imaginada del ser humano a lo que se refiere en primera instancia el problema mente-cuerpo. Sin embargo, podemos distinguir tres formulaciones distintas de tal problema, las cuales inciden en un aspecto u otro del mismo según el contexto de su propia formulación. Tales formulaciones son la religiosa, la filosófica y la que

consideramos propiamente científica. La formulación religiosa del problema mente-cuerpo entiende que la dualidad es entre cuerpo y alma, es decir, entre lo que tenemos de común con las realidades materiales de *este* mundo y nuestra alma o espíritu en cuanto realidad inmaterial y ajena a *este* mundo. Llamaré a esta formulación del problema "el problema de San Agustín", ya que este santo experimentó como pocas personas el enfrentamiento doloroso entre las pasiones de su cuerpo y el anhelo de inmortalidad y paz de su alma.

La formulación filosófica del problema mente-cuerpo entiende que la dualidad se da entre el cuerpo, que es público, observable por los demás y sometido a las leyes mecánicas, y la mente o pensamiento, que es una realidad privada, observable sólo por ella misma y en apariencia no sometida a las leyes mecánicas. Llamaré a esta formulación del problema "el problema de Descartes", ya que, como vimos en el capítulo 2, fue René Descartes quien, al haber defendido la radical diferencia entre extensión corporal y pensamiento inextenso, dejó planteadas las dificultades para explicar las relaciones que, por otra parte, todos podemos experimentar entre cuerpo y mente. Tal es el tradicional problema mente-cuerpo o, si se prefiere, el problema mente-cuerpo por antonomasia.

Sin embargo, existe una tercera formulación del problema mente-cuerpo, que es su formulación científica actual. Para tal formulación la dualidad se da, o se daría, por un lado, entre el cerebro (o bien el sistema nervioso central), en cuanto centro que realiza las funciones de recibir los estímulos del ambiente, procesarlos de acuerdo con la experiencia acumulada y con ciertas estructuras y, por último, poner en marcha las respuestas adecuadas, y, por otro lado, la mente, en cuanto los procesos mismos de recepción de información, de procesamiento de la información recibida y de ejecución o inhibición de respuestas. Tal es el problema mente-cerebro, en el que se discute si los procesos mentales (tal como los estudia la psicología cognitiva o la inteligencia artificial) se identifican o no con los procesos cerebrales (tal como los estudia la neurociencia), y en caso de no identificarse cuáles son las relaciones entre unos y otros. Llamaré a esta nueva formula-

ción del problema "el problema de Penfield", ya que fue Wilder Penfield (1891-1976), célebre neurocirujano americano, uno de los científicos más destacados en el estudio de las relaciones entre mente y cerebro, en particular en su libro *The Mystery of the Mind* (1975).

De las tres formulaciones del problema mente-cuerpo, nos interesa el problema de Penfield o problema mente-cerebro. El problema de San Agustín, aunque perfectamente legítimo e importante, es una cuestión teológica en la que no pretendo entrar. Con todo, las soluciones al problema de Penfield no deberán ser indiferentes para el problema cuerpo-alma. En cuanto al problema de Descartes, ya pusimos de relieve sus presupuestos erróneos, es decir, que la mente debe concebirse como una substancia, que una realidad extensa no puede pensar y, por último, su rechazo de las mentes mecánicas. Con todo, al analizar la solución de John Eccles al problema mente-cerebro veremos reaparecer de algún modo el problema mente-cuerpo.

2. Las soluciones al problema mente-cerebro

Respecto del problema de Penfield o problema mente-cerebro, podemos enumerar distintas posturas actuales. En primer lugar, y puesto que la psicología conductista aún está viva y coleando, debemos anotar el conductismo, que estudiamos en el capítulo V, como postura que sostiene que el problema no es un auténtico problema y ni siquiera debe plantearse. En segundo lugar, debemos anotar el materialismo o teoría de la identidad mente-cerebro, que estudiamos en el capítulo VII, para el cual el problema puede plantearse pero queda disuelto tras su consideración. En tercer lugar, tenemos el funcionalismo, que también consideramos en el capítulo VII, que significa una especie de "tiempo muerto" en el problema, puesto que defiende que los procesos mentales son independientes de su eventual soporte físico o de los órganos específicos que los produzcan. Finalmente, podemos registrar las soluciones fieles al planteamiento del problema, esto es, que mantienen algún tipo o grado de dualismo entre mente y cerebro, ya que

sin la admisión de algún tipo o grado de dualismo entre mente y cerebro no existe propiamente el problema mente-cerebro. Tales soluciones son el emergentismo, el dualismo radical y lo que denominé en mi trabajo "Procesos mentales y cognitivismo" (1992) dualismo razonable.

Estudiaremos con cierto detalle el emergentismo, tal como aparece en John Searle, y el dualismo radical, tal como es defendido por John Eccles, y me referiré, al criticar la posición de Eccles a mi dualismo razonable. De momento, recordaremos algunas ideas propias de las otras soluciones, esto es, el conductismo, el materialismo y el funcionalismo.

Para el conductismo el problema de Penfield es un pseudoproblema, ya que para esta postura la mente no existe propiamente y el cerebro es irrelevante para el estudio de los fenómenos psicológicos. El conductismo propugna el abandono del lenguaje mentalista (términos como deseo, creencia, sentimiento o propósito) en las explicaciones psicológicas, donde sólo debe estar presente el lenguaje relativo a la descripción de estímulos y respuestas así como a las leyes que gobiernan la conducta. La mente, como resulta obvio, es un concepto mentalista (que refleja una supuesta interioridad del ser humano) y por ende es ajena a una postura filosófica y científica centrada en la exterioridad del hombre, en su ambiente y en sus respuestas públicamente observables. Por otra parte, los conductistas llevan su insistencia en la exterioridad de la conducta hasta tal punto que menosprecian o bien descuidan el estudio de los procesos cerebrales, sospechoso de llevarnos a algún tipo de retorno mentalista. En consecuencia, puesto que no existe la mente ni importa el cerebro, no tiene sentido hablar de un problema mente-cerebro.

En cuanto al materialismo, se plantea la posibilidad de distinguir entre mente y cerebro, o mejor, entre procesos mentales y procesos cerebrales, pero sostiene que una consideración de ambas clases de procesos lleva de modo contundente a la identificación entre una y otra, en el específico sentido de que cualquier tipo de proceso mental es idéntico a un tipo determinado de proceso cerebral. El materialismo o teoría de la identidad mente-cerebro, aunque considera legí-

timo el problema de Penfield, procede a su eliminación tras dar su solución. Esta doctrina, como señalé en su momento, tiene un indudable atractivo, ya que promete (aunque no ha cumplido) la eliminación de la psicología y con ello una simplificación en el cuerpo general de la ciencia. Pero además también promete sustituir las especulaciones e hipótesis a las que a veces se ven obligados los psicólogos por los estudios concretos y positivos del neurocientífico (si en verdad siempre tienen tal carácter). Pero, como ya vimos, la teoría de la identidad mente-cerebro se enfrenta a lo que denominé argumento antimaterialista de las máquinas, es decir, a la evidencia empírica de que algunas máquinas, carentes de procesos cerebrales, tienen no obstante procesos mentales.

Respecto del funcionalismo debemos recordar que constituye una teoría general de los procesos mentales que goza de gran prestigio y de aceptación generalizada entre los filósofos, los psicólogos cognitivistas y los científicos de la inteligencia artificial. Para el funcionalismo, los procesos mentales son procesos internos con un papel o rol causal sobre la conducta, que constituyen *funciones* mediadoras entre entradas sensoriales y salidas motoras, y que se definen y caracterizan con total independencia de su soporte físico. Como dice Putnam, en *Reason, Truth and History* (1981), la misma propiedad psicológica (por ejemplo, estar iracundo) puede ser una propiedad perteneciente a miembros de miles de especies diferentes, que pueden poseer una composición física o química completamente distinta (algunas de estas especies podrían ser extraterrestres y quizás algún día los robots muestren ira).

Tal teoría es, en efecto, general, ya que puede ser aceptada por el psicólogo cognitivo (para quien el soporte de los procesos mentales puede ser el cerebro), también puede ser aceptada por el científico de inteligencia artificial (para quien el soporte de los procesos mentales son los circuitos electrónicos), por el filósofo de orientación materialista (para quien el soporte de los procesos mentales es el cerebro) e incluso por el filósofo de orientación espiritualista (para quien algunos o todos los procesos mentales tendrían su soporte

último en el espíritu). Por ello he calificado al funcionalismo de "tiempo muerto" en la discusión mente-cerebro, ya que el problema de Penfield queda en suspenso o aplazado con el funcionalismo, en espera de mejor momento para su discusión o en espera de nuevas evidencias empíricas.

3. El emergentismo de Searle

John Searle, de cuya célebre habitación china nos hemos ocupado en el capítulo anterior, defiende, en su libro *Minds, Brains and Science* (1984), la tesis general de que la mente y el cuerpo interactúan aunque no son dos cosas diferentes, puesto que los fenómenos mentales son solamente rasgos del cerebro. Para Searle, tal postura es a la vez una aserción de fisicalismo y de mentalismo, esto es, podemos añadir, se trata de defender al mismo tiempo que los procesos mentales no son independientes de los procesos físicos (en concreto, de los procesos cerebrales) y que los procesos mentales no se reducen a los procesos cerebrales tal como los estudia el neurocientífico.

Esta posición puede denominarse emergentismo, ya que, cuando se sostiene que los procesos mentales no pueden darse sin procesos cerebrales pero tampoco se reducen a éstos, los procesos mentales aparecen como fenómenos o propiedades que emergen de los procesos cerebrales. El propio Searle, en su trabajo "Minds and Brains Without Programs" (1987) acepta el término emergentismo para indicar su postura de que las propiedades mentales son propiedades emergentes de los sistemas neurofisiológicos, aunque no acepta tal término para indicar su postura si con tal término se defiende que las propiedades mentales emergen mediante algún proceso no-físico de carácter misterioso.

Para nuestro autor, tal como señala en *Minds, Brains and Science,* hay cuatro rasgos de los fenómenos mentales (tomados globalmente) difíciles de encajar en la concepción del mundo como compuesto de cosas materiales. El más importante de tales rasgos es la conciencia (aunque no todos los procesos mentales sean conscientes), en el sentido, podemos añadir, de darnos cuenta de lo que hacemos. Para Searle,

la conciencia es el hecho central de la existencia específicamente humana y, sin embargo, es difícil ver cómo sistemas meramente físicos pueden tener conciencia. El segundo rasgo de la mente difícil de encajar en una concepción materialista del mundo es la intencionalidad (aunque tampoco todos los procesos mentales son intencionales), es decir, el rasgo mediante el cual nuestros estados mentales se dirigen o refieren a objetos o estados de cosas del mundo distintos de los propios estados mentales. Nuestros deseos, esperanzas, temores, creencias o intenciones (entendidas éstas como propósitos) tienen intencionalidad en cuanto se refieren al mundo distinto de la mente y, sin embargo, es difícil concebir que una realidad meramente material pueda representar algo.

El tercer rasgo de la mente difícil de acomodar en una concepción materialista de la realidad es la subjetividad de los estados mentales (y ahora podemos decir que todos los procesos mentales son subjetivos). Tal subjetividad indica el hecho de que cada cual puede conocer sus estados mentales internos mientras que los demás no pueden conocerlos (directamente). Esto es, se trata de lo que denominamos el privilegio del acceso directo de cada sujeto humano a sus propios procesos mentales. Sin embargo, parece difícil acomodar este carácter subjetivo de los estados mentales con el carácter objetivo de la concepción científica materialista. Finalmente, Searle señala un cuarto problema, el problema de la causación mental. Nuestros pensamientos y sensaciones tienen de hecho algún efecto causal sobre el mundo físico; por ejemplo, decido levantar mi brazo y mi brazo se levanta. En suma, nuestros procesos mentales (aunque no todos) causan efectos físicos. Y sin embargo, no se comprende fácilmente que algo mental pueda tener influencia física.

La respuesta de John Searle a estas dificultades consiste en sostener lo que podríamos calificar de un cierto dualismo entre mente y cerebro, pero un dualismo que no admite que mente y cerebro sean dos cosas distintas, sino que los procesos mentales en cuanto tales son aspectos o rasgos diferenciables de los procesos neurofisiológicos aunque no independientes de éstos. En concreto, la tesis (emergentista) de Searle es que

todos los fenómenos mentales, ya sean conscientes o inconscientes, están causados por procesos cerebrales (y en esta medida emergen de ellos) y al mismo tiempo están realizados en los procesos cerebrales (y en esta medida no son radicalmente distintos de ellos).

Para aclarar su tesis, nuestro autor recurre a la distinción entre micropropiedades y macropropiedades de los sistemas físicos. Si, por ejemplo, consideramos una mesa o un vaso de agua, cada uno de estos objetos está compuesto de micropartículas, que tienen rasgos al nivel de moléculas y átomos; tales son sus micropropiedades. Pero además, cada objeto tiene rasgos, como la solidez de la mesa, la liquidez del agua o la transparencia del vaso, que son propiedades superficiales o globales; tales son sus macropropiedades. Por otra parte, ocurre que muchas de esas propiedades superficiales o globales pueden explicarse causalmente por la conducta de los elementos del micronivel; por ejemplo, la solidez de la mesa se explica por la estructura de enrejado ocupada por las moléculas de las que se compone. Es decir, hay macropropiedades que se explican causalmente por la conducta de los elementos del micronivel. De modo más claro, el rasgo superficial o macropropiedad es causado por la conducta de los microelementos, pero al mismo tiempo está realizado en el sistema que está compuesto de los microelementos.

Podemos, según Searle, aplicar esta distinción al estudio de la mente para dar cuenta de las relaciones de la mente con el cerebro. Lo mismo que la solidez de la mesa es causada por la conducta de elementos del micronivel y, con todo, es al mismo tiempo un rasgo realizado en el sistema de microelementos, así exactamente los fenómenos mentales son causados por procesos que tienen lugar en el cerebro en el micronivel neuronal y, al mismo tiempo, son rasgos que están realizados en el sistema neuronal. Para nuestro autor, nada hay más común en la naturaleza que el que rasgos superficiales (macropropiedades) sean a la vez causados por y realizados en una microestructura, y tal es la relación entre mente y cerebro. En el cerebro hay dos niveles de descripción causalmente reales: el macronivel de procesos mentales y el micronivel de procesos

neuronales. Podemos resumir la tesis de Searle diciendo que los procesos mentales están causados por los procesos cerebrales y al tiempo están realizados en los procesos cerebrales.

En su trabajo de 1987 "Minds and Brains Without Programs", Searle insiste en la distinción entre macropropiedades y micropropiedades, y señala que los rasgos intrínsecamente mentales del universo (conciencia, intencionalidad, subjetividad y causación intencional) son justamente rasgos físicos de nivel superior de los cerebros. Pero las propiedades mentales se caracterizan explícitamente como propiedades emergentes de los sistemas neurofisiológicos. Asimismo, el axioma primero de las conclusiones de este artículo dice tajantemente que los cerebros causan las mentes. En suma, podemos comentar, para Searle los procesos mentales son algo diferenciado de los procesos cerebrales en cuanto neuronales, aunque no existan al margen de los procesos neuronales; de esta manera podrá distinguirse entre los procesos mentales tal como los estudia la psicología y los procesos cerebrales tal como los estudia la neurociencia.

Por último, en su libro *The Rediscovery of the Mind* (1992), que dedica de modo preferente al estudio de la conciencia, Searle critica tanto el dualismo como el materialismo, reafirmando su posición emergentista. Para él, la conciencia es un rasgo irreducible de la realidad física o, más precisamente, es un rasgo emergente de ciertos sistemas de neuronas, de la misma manera que la solidez y la liquidez son rasgos emergentes de sistemas de moléculas.

El emergentismo de Searle tiene un aspecto positivo y otro negativo. Por un lado, al no sucumbir al reduccionismo mantiene la autonomía de la psicología, la cual, firmemente apoyada en la neurociencia, no se confunde ni reduce a la neurociencia. Con ello, la explicación psicológica sigue teniendo su dominio propio de aplicación y la mente es un tema que debe reabrirse o redescubrirse. Pero, por otro lado, la dependencia de los procesos mentales exclusivamente de los procesos cerebrales lleva a Searle a excluir la posibilidad de mentes mecánicas, frente a la evidencia empírica de la existencia de máquinas que piensan. El biologismo de nuestro autor le

impide aceptar que los procesos mentales pueden tener una base física distinta de los sistemas neuronales.

4. El dualismo de Eccles y su trasfondo metafísico

El neurólogo John Eccles (nacido en Australia en 1903), Premio Nobel de Medicina de 1963, ha defendido en nuestros días un dualismo radical interaccionista entre la mente y el cerebro. Su hipótesis general es que la mente autoconsciente es distinta del cerebro y ejerce una función superior de interpretación y control de los procesos neuronales, en virtud de la interacción entre mente y cerebro.

Inicialmente, en su libro *The Neurophysiological Basis of Mind: The Principles of Neurophysiology* (1953), Eccles, siguiendo al gran neurólogo Sherrington, había postulado la existencia de un área especial del cerebro en enlace con la conciencia. Posteriormente, tras los trabajos de Roger Sperry sobre comisurotomía (o corte del cuerpo calloso que une ambos hemisferios cerebrales), y tras la presentación por Karl Popper (nacido en Viena en 1902) de su teoría de los tres mundos, Eccles elaboró su hipótesis dualista. Tal hipótesis ya aparece en su trabajo "Cerebral Activity and Consciousness" (1974), pero está desarrollada en detalle en el libro que escribieron Popper y el propio Eccles titulado *The Self and its Brain* (1977).

Más adelante analizaremos los experimentos de Roger Sperry sobre pacientes con comisurotomía, así como la interpretación que hace Eccles de los resultados de tales experimentos. De momento nos centraremos en la exposición de la teoría de Popper de los tres mundos.

El filósofo Karl Popper presentó, en el III Congreso Internacional de Lógica, Metodología y Filosofía de la Ciencia (Amsterdam, 1967), una ponencia sobre su teoría de los tres mundos, publicada por primera vez en las Actas del Congreso (1968) y luego incorporada a su libro *Objective Knowledge* (1972). Según esta teoría, todo lo que existe y toda nuestra experiencia está contenida en uno de estos mundos: 1) Mundo 1, que está constituido por los objetos y estados físicos (objetos

naturales, inanimados o biológicos, incluido el cerebro, así como los objetos artificiales), 2) Mundo 2, que está constituido por las conciencias y los conocimientos subjetivos (percepciones, sentimientos, intenciones, recuerdos, etc.), y 3) Mundo 3, que es el mundo de la cultura creada por el hombre, incluidos todos los conocimientos objetivos (problemas científicos, argumentos críticos y sistemas teóricos). Esta teoría supone distinguir dos sentidos de conocimiento: conocimiento en sentido subjetivo, que consiste en un estado mental o de conciencia, y conocimiento en sentido objetivo, que consiste en problemas, teorías y argumentos en cuanto tales.

Además, según Popper, se dan interacciones entre los tres mundos. Hay interacciones entre el mundo 2 (mundo de las conciencias) y el mundo 3 (mundo de los productos culturales), ya que el mundo 3 es una creación humana, aunque, según nuestro autor, una vez creado es en gran medida autónomo. A su vez, también hay interacciones entre el mundo 1 (mundo de los objetos físicos) y el mundo 3, puesto que algunas realidades físicas, como los aviones, son también productos culturales. Finalmente, y es lo que más interesa a nuestro problema de la relación mente-cerebro, existen interacciones entre el mundo 1 (donde se incluye el cerebro) y el mundo 2 (mundo de los estados mentales); la relación del mundo 1 al mundo 2 se advierte, por ejemplo, en la percepción consciente, mientras que la relación del mundo 2 al mundo 1 se advierte, por ejemplo, en el acto voluntario.

Eccles acepta toda esta teoría de los tres mundos, añadiendo que el mundo 2 tiene tres componentes principales: 1) el sentido externo, o percepciones producidas por los datos de los órganos sensoriales, 2) el sentido interno, que comprende pensamientos, recuerdos, intenciones, representaciones, emociones o sentimientos, y 3) el yo, que está situado en el núcleo del mundo 2 y que es la base de la identidad y continuidad personal que experimentamos. Ninguno de estos componentes, según Eccles, existe como tal en el mundo 1, mientras que el cerebro se da en el mundo 1.

Este trasfondo metafísico del dualismo de John Eccles aparece en todas las obras en que defiende el dualismo mente-

cerebro. No sólo en *The Self and its Brain*, libro escrito, como dijimos, en colaboración con Popper, sino en sus libros posteriores, como *The Human Psyche* (1980), *Gehirn und Geist* (1980) escrito con la colaboración del biólogo suizo Hans Zeier, o *Evolution of the Brain: Creation of the Self* (1989).

5. Fundamentos neurológicos del dualismo de Eccles

Tal como señala Hans Zeier, en el libro que acabamos de citar, el neocortex o corteza cerebral constituye la más importante conquista de los mamíferos. En los seres humanos, el neocortex es tan extenso que cubre completamente las otras porciones del cerebro y además se presenta muy plegado con lo que aumenta de modo considerable su superficie. Desde el punto de vista funcional, la corteza cerebral se divide en tres tipos de áreas: áreas motoras, áreas sensoriales y áreas de asociación. La estimulación eléctrica de las áreas motoras provoca contracciones en los músculos correspondientes, mientras que la estimulación eléctrica de las áreas sensoriales produce las sensaciones correspondientes. Ambos tipos de áreas están subordinadas a las áreas de asociación, las cuales no poseen conexión directa con la periferia, siendo su misión principal la de relacionar entre sí las diversas áreas de la corteza cerebral. La corteza de asociación es la sede de las funciones de integración y coordinación.

A su vez, el propio Eccles, también en *Gehirn und Geist* (así como en otras obras suyas), se refiere a los estudios de Janos Szentágothai, tal como se exponen en su "The Basic Neuronal Circuit of the Neocortex" (1972), sobre la microestructura del neocortex. Este autor tras analizar la estructura de áreas sensoriales específicas generaliza sus resultados a toda la corteza cerebral. En resumen, para Szentágothai, el neocortex está compuesto por columnas verticales con una sección media de 0'2 mm^2 y unas diez mil neuronas por columna. Tales columnas, llamadas módulos, son la unidad básica de la corteza cerebral, tanto en el plano estructural como en el plano funcional, siendo comparables a los micro-

circuitos integrados de la electrónica. Los módulos, o circuitos neuronales básicos, tienen canales de entrada (fibras aferentes), conexiones neuronales complejas en el interior del módulo, y canales de salida (básicamente axones de las neuronas piramidales). Por otra parte, entre los distintos módulos se producen excitaciones de diferentes intensidades así como también inhibiciones, además de circuitos de retorno con acción sostenida.

Sin embargo, el fundamento neurológico inmediato del dualismo de Eccles, tal como se expone también en varias de sus obras, se encuentra en el distinto carácter de los dos hemisferios cerebrales así como en los experimentos de comisurotomía realizados por Sperry. En efecto, es habitual distinguir entre hemisferio cerebral izquierdo y hemisferio cerebral derecho en cuanto, en la mayoría de los individuos humanos, desempeñan papeles distintos. El hemisferio izquierdo o dominante posee las facultades verbales e ideacionales, siendo además el hemisferio analítico y aritmético. En cambio, el hemisferio derecho o subordinado carece de capacidad verbal o lingüística y posee escasa capacidad ideacional, aunque posee las capacidades pictórica y musical, siendo además el hemisferio sintético y geométrico.

A partir de los años sesenta, Roger Sperry y sus colaboradores realizaron en el Instituto de Tecnología de California (en Pasadena) una serie de experimentos con pacientes a los que se había practicado la operación de comisurotomía. Tal operación consiste en el corte total del cuerpo calloso, es decir, del paquete de unos doscientos millones de fibras nerviosas que unen los dos hemisferios cerebrales asegurando así la cooperación entre ambos. La comisurotomía se realiza en pacientes epilépticos muy graves, consiguiéndose una notable mejoría. Esta operación se había realizado en varias ocasiones no advirtiéndose nada importante en la conducta del paciente, ya que las actividades del cuerpo de mantenerse en pie, pasear, nadar, dormir o despertarse continuaban coordinadas merced a las uniones de los hemisferios con las regiones del cerebro inferior, que no habían sido afectadas por el corte del cuerpo calloso; además la unión de ambos hemisferios

anterior a la operación garantizaba en cada hemisferio el recuerdo de esos años de ejecución conjunta. Sólo los estudios detallados de Sperry, Gazzaniga y Bogen, publicados en "Interhemispheric relationships: the neocortical commissures; syndromes of hemisphere disconnection" (1969), revelaron notables alteraciones en los pacientes.

En efecto, se proporcionaba información visual a los sujetos con el cerebro dividido, o bien a su campo visual derecho, captado por el hemisferio izquierdo, o bien a su campo visual izquierdo, captado por el hemisferio derecho. (Este cruce de la percepción visual es habitual en todo cerebro). Los sujetos tenían, como es corriente, sus áreas de lenguaje en el hemisferio izquierdo. El resultado fue que las actividades neuronales en el hemisferio derecho, en concreto sus percepciones visuales, no eran comunicadas verbalmente por el sujeto, mientras que las percepciones visuales en el hemisferio izquierdo eran comunicadas verbalmente por el sujeto. Es decir, el sujeto sólo comunicaba mediante el lenguaje las percepciones recibidas en el hemisferio izquierdo, donde se encuentran las áreas de lenguaje, procedentes del campo visual derecho.

La interpretación que hace Eccles de estos experimentos es que la mente autoconsciente se relaciona con el hemisferio izquierdo, pero no (o bien sólo de modo escaso y ocasional) con el hemisferio derecho. En general, pues, las experiencias conscientes del ser humano surgen sólo en relación con las actividades neuronales del hemisferio izquierdo o dominante.

6. La hipótesis dualista interaccionista

Tomando como base esos datos y experimentos neurológicos, y utilizando la teoría de los tres mundos popperiana, John Eccles formula su hipótesis dualista radical interaccionista, esto es, mente y cerebro son dos entidades distintas (la una perteneciente al mundo 2 y la otra al mundo 1) aunque entre ambas existen interacciones.

Más en detalle, y tal como dice en *Gehirn und Geist*, cabe establecer los seis principios siguientes. En primer lugar, las

experiencias de la mente autoconsciente están en relación con los procesos neuronales que tienen lugar en las áreas asociativas del neocortex; tal relación es una interacción, produciendo, por tanto, cierto grado de correspondencia pero no identidad (frente, podemos decir, al materialismo). En segundo lugar, la experiencia de la mente autoconsciente tiene un carácter unitario, y además podemos concentrarnos en éste o aquél aspecto de la actividad cerebral (fenómeno de atención). Asimismo, las informaciones procedentes de los órganos sensoriales son transmitidas al cerebro, pero, al cruzar la frontera entre la corteza cerebral asociativa y la mente, se transforman milagrosamente en las variadas experiencias de nuestro mundo perceptivo, que son distintas de los procesos cerebrales.

En cuarto lugar, la mente autoconsciente es capaz de actuar sobre los procesos cerebrales; esto resulta obvio en los actos voluntarios, pero también durante nuestro estado de vigilia desencadenamos procesos en la maquinaria neuronal como, por ejemplo, al recordar algo, hacer un cálculo mental o buscar la frase apropiada para expresar una idea. En quinto lugar, la interacción a través de la frontera entre la mente y la corteza cerebral asociativa sólo tiene lugar cuando existe un elevado nivel de actividad diversificada en el mecanismo neuronal del cerebro asociativo; cuando el nivel de actividad es muy bajo, como, por ejemplo, en la anestesia o en estado de coma, el sujeto se halla inconsciente. Finalmente, debemos establecer que el mundo 1 (mundo de los objetos físicos) es incompleto, pudiendo ser influido por el mundo 2 (mundo de los estados mentales) y a su vez actuar sobre él; esto es, el mundo 1 está abierto causalmente al mundo 2.

Sobre estos principios, Eccles formula de modo más completo su hipótesis dualista. Según tal hipótesis, la mente autoconsciente se dedica activamente a leer los mensajes de multitud de centros con el más alto nivel de actividad cerebral, a saber, las zonas asociativas del hemisferio cerebral dominante o izquierdo. La mente autoconsciente escoge entre las informaciones de estos centros según la orientación de su atención, y las integra en un todo. Por otra parte, la mente autoconsciente actúa sobre esos centros nerviosos, modifican-

do las pautas espaciotemporales dinámicas de los procesos neuronales. En suma, la mente autoconsciente ejerce una función superior de interpretación y control de los procesos neuronales. A su vez, la unidad de la experiencia consciente la produce la mente y no la maquinaria neuronal de la corteza asociativa del cerebro.

De manera aún más concreta, según Eccles, la mente explora y también modifica las actividades de los módulos de la corteza cerebral asociativa. Por tanto, son los módulos, con su compleja organización e intensa actividad, los componentes del mundo físico abiertos al mundo de la mente, tanto en el sentido de dar como de recibir. Ahora bien, no todos los módulos, añade Eccles, poseen esta propiedad de estar abiertos al mundo de la mente, sino solamente los módulos del cerebro asociativo con un adecuado nivel de actividad.

La acción de la mente autoconsciente (que, para nuestro autor, es una entidad autosubsistente ajena al mundo físico) sobre los procesos neuronales parece difícil de aceptar. Sin embargo, Eccles se refiere a los experimentos de Benjamin Libet y de Hans Helmut Kornhuber para poner de relieve el papel activo de la mente autoconsciente sobre los procesos cerebrales.

Por un lado, Benjamin Libet, tal como aparece publicado en su trabajo "Electrical stimulation of cortex in human subjects, and conscious sensory aspects" (1973), realizó experimentos sobre el tiempo en la percepción consciente. En ellos se manifestó que entre la estimulación eléctrica directa repetida de la corteza cerebral somatosensorial y la aparición de la percepción consciente correspondiente, tal como informaba el sujeto, había un intervalo temporal de 0'5 segundos en caso de estimulaciones débiles. Si la estimulación era periférica, en la piel, y también débil, podría esperarse que su percepción consciente fuese más tardía y, sin embargo, el sujeto la experimentaba antes que la estimulación correspondiente directa en la corteza cerebral. Eccles interpreta estos resultados diciendo que esta anticipación de la percepción consciente es atribuible a la capacidad de la mente autoconsciente para manipular el tiempo neuronal.

Por otro lado, Hans Helmut Kornhuber, en su trabajo "Cerebral cortex, cerebellum and basal ganglia: An introduction to their motor functions" (1974), publicó los resultados de sus experimentos sobre la acción voluntaria. El sujeto de experimentación hacía voluntariamente una flexión rápida del dedo índice derecho, a intervalos irregulares de varios segundos. Se trazaron las curvas medias correspondientes a doscientos cincuenta registros de potenciales eléctricos en varias zonas de la superficie craneal. Lo sorprendente es que se encontró un potencial eléctrico generado en la corteza cerebral antes de la ejecución de la acción voluntaria; tal potencial eléctrico, denominado "potencial de alerta", comenzaba casi 800 milisegundos antes del inicio de los potenciales de acción musculares. La interpretación de Eccles es que la mente autoconsciente constituye y aprovecha tal potencial de alerta para configurar las pautas de actividad de los módulos de la corteza cerebral.

7. Evaluación del dualismo de Eccles

El dualismo radical mente-cerebro interaccionista defendido por John Eccles tiene el atractivo inicial de estar aparentemente apoyado en datos y experimentos neurológicos que son utilizados por un neurólogo de indudable prestigio. Sin embargo, cabe tener dudas sobre si la interpretación de Eccles de tales experimentos es la correcta.

Para empezar, Roger Sperry, en "Mind-Brain Interaction: mentalism, yes; dualism, no" (1980), desautoriza la interpretación dualista que Eccles (y Popper) hace de sus trabajos e hipótesis neurológicas. Para Sperry, los fenómenos conscientes son propiedades funcionales emergentes del procesamiento cerebral y, al tiempo, ejercen un papel de control activo como determinantes causales en la configuración de los patrones de flujo de la excitación cerebral. Una vez generados a partir de los eventos neuronales, los patrones y programas mentales tienen sus propias cualidades subjetivas y (lo que es más importante) operan e interactúan de acuerdo con sus propias leyes y principios causales, que son diferentes y no pueden reducirse a los de la neurofisiología.

Para este autor, las entidades mentales transcienden lo fisiológico al igual que lo fisiológico transciende lo molecular y lo molecular lo atómico. Esta idea (que podemos calificar de emergentismo), señala Sperry, es muy diferente de las de intervención fantasmal extrafísica defendidas por Eccles y Popper. Más aún, esta idea de la relación entre mente y cerebro, añade nuestro autor, no sólo refuta las doctrinas del conductismo y del materialismo, como Popper y Eccles infieren con razón, sino que también con igual fuerza descarta el dualismo. En suma, podemos añadir, el propio Sperry, en cuyos trabajos se ha basado Eccles para concluir su dualismo, rechaza tal interpretación dualista de sus trabajos defendiendo, por su parte, un emergentismo de los procesos mentales, que no es un dualismo radical sino, a lo sumo, un dualismo mitigado.

Por otra parte, los experimentos de Benjamin Libet acerca del tiempo neuronal (en cuanto distinto aparentemente del tiempo físico) y los experimentos de Kornhuber acerca del potencial de alerta (en cuanto anterior a los potenciales de acción musculares) son argumentos de peso a favor de un factor no-físico en la percepción consciente y en la acción voluntaria respectivamente. Sin embargo, Roger Penrose, en su libro *The Emperor's New Mind* (1989), hace la sugerencia de que es probable que estemos equivocados al aplicar las reglas físicas usuales para el tiempo cuando consideramos la conciencia. Con todo, me parece que la sugerencia de Penrose puede llevarnos en dos direcciones distintas. Por un lado, podemos pensar que el tiempo de la conciencia tiene un carácter especial que, no obstante, no rebasa las coordenadas y pautas físicas (aunque sea *otra* ciencia física). Por otro lado, podemos creer que el tiempo de la conciencia humana no es reducible a categoría física alguna.

De todas maneras, el mayor problema del dualismo de Eccles está en su formulación en términos de la teoría de los tres mundos de Popper. En efecto, tal formulación supone una clara petición de principio, ya que, de entrada, el cerebro se incluye en el mundo1 (mundo de los objetos físicos), lo cual no suscita recelo alguno, mientras que la mente se declara

constituyendo el mundo 2, siendo, por definición, distintos ambos mundos. A partir de aquí ya está decidido todo, y los argumentos son simples adornos añadidos a una tarta prefabricada.

Otra seria dificultad del dualismo de Eccles es que considera, por ejemplo en *The Human Psyche* (1980), que el dualismo cerebro-mente debe transformarse en el dualismo cerebro-psique y asimismo que la psique es el alma, objeto de creación divina. Según ello, los procesos mentales serían espirituales y entonces nuestro autor se vería obligado a admitir que los animales también tienen alma o espíritu, ya que es evidente que los animales tienen procesos mentales. El propio Eccles parece responder a esta objeción cuando en su libro *Evolution of the Brain: Creation of the Self* (1989) distingue entre el mundo 2 de un mamífero y el mundo 2 de un ser humano (figuras 8.5 y 10.4). En efecto, el mundo 2 de un mamífero consta de sentido interno y de sentido externo, careciendo del componente central que es la psique o alma, mientras que el mundo 2 de un ser humano posee los tres componentes (que analizamos anteriormente).

Cuando discutimos el materialismo, señalamos la existencia en los seres humanos de distintos tipos de procesos mentales que no parecen poder admitir una explicación en términos puramente físico-químicos y que, por tanto, podrían calificarse de procesos no-físicos o incluso espirituales. Tales tipos de procesos mentales eran los fenómenos de autoconciencia (en el sentido de conciencia de nuestra identidad personal), los fenómenos de formación de un proyecto vital personal, y las voliciones libres o indeterministas.

Podemos admitir, y esta es mi creencia, que estos procesos mentales humanos son no-físicos, pero no podemos admitir con Eccles que todos los procesos mentales humanos son no-físicos. Es decir, frente a la hipótesis de Eccles, los procesos mentales humanos no son globalmente no-físicos ni pertenecen todos ellos al mundo 2, sino que algunos son no-físicos mientras que la inmensa mayoría son físicos, perteneciendo por tanto al mundo 1 popperiano.

La hipótesis de John Eccles fracasa como teoría general

de la mente. Por una parte, no explica adecuadamente la mente animal, ya que, si su hipótesis se aplica a los animales, resulta muy aventurado atribuir a los animales una mente autoconsciente de carácter no-físico. Por otra parte, no explica en absoluto la mente mecánica, puesto que los procesos mentales que se dan en algunas máquinas (como vimos en el capítulo anterior) no tienen interacción alguna con procesos cerebrales.

Mi propia hipótesis, que califico de dualismo razonable, defiende la existencia de tres categorías generales de procesos mentales. En primer lugar, procesos mentales-cerebrales (y por tanto físicos), que se dan tanto en los seres humanos como en los animales (en particular en los vertebrados). En segundo lugar, procesos mentales-físicos (aunque no cerebrales), que se presentan en algunas máquinas. Y en tercer lugar, procesos mentales no-físicos (y probablemente espirituales), que se dan exclusivamente en los seres humanos. Por otra parte, respecto de las dos primeras categorías debe defenderse una postura emergentista; los procesos mentales-cerebrales son propiedades funcionales emergentes de los procesos neuronales, mientras que los procesos mentales-físicos son propiedades funcionales emergentes de los procesos físicos de computación. Tal emergentismo supone un dualismo mitigado, ya que los procesos mentales de ambas categorías no se reducen a sus soportes físicos. En cambio, respecto de los procesos mentales no-físicos puede defenderse una postura dualista más fuerte, en el sentido de que estos procesos mentales aparecen desarrollando una causación no-física independiente de los procesos neuronales y al mismo tiempo utilizando estos procesos neuronales.

10

Mentalismo y psicología cognitiva

1. Origen y desarrollo de la psicología cognitiva

El origen de la psicología cognitiva, a la que ya nos hemos referido anteriormente en varias ocasiones, está unido al surgimiento de las ciencias cognitivas. A su vez, las ciencias cognitivas constituyen un campo de investigación interdisciplinar cuyo tema aglutinante y central es el estudio del fenómeno de la cognición en seres humanos, animales y máquinas. No podemos pretender aquí una caracterización detallada de la cognición, pero podemos indicar que, en primera instancia, puede entenderse como sinónimo de conocimiento, en cuanto recepción de información, aunque, de modo más preciso, debe entenderse también como uso y manejo de conocimiento, es decir, como manipulación de información.

Dentro del campo de las ciencias cognitivas, las ciencias básicas son la psicología cognitiva y la inteligencia artificial, mientras que la lógica, la informática, la neurociencia y la lingüística son ciencias de carácter instrumental. Por último, la filosofía de la mente, debidamente entendida, puede desempeñar el papel de metaciencia cognitiva o filosofía de las ciencias cognitivas.

Podemos aceptar la idea de Howard Gardner, expuesta en su libro *The Mind's New Science. A History of the Cognitive Revolution* (1985), de que las ciencias cognitivas se abren camino en 1948 en el simposio celebrado en Pasadena (Cali-

fornia) patrocinado por la Fundación Hixon. En tal simposio, sobre "Los mecanismos cerebrales en la conducta", ocurrieron cosas importantes y novedosas. Por ejemplo, el matemático y lógico John von Neumann (1903-1957) habló de la analogía entre el computador y el cerebro, el neurocientífico Warren McCulloch (1898-1969) explicó el modo en que el cerebro procesa información, y, entre otras novedades, el psicólogo Karl Lashley (1890-1958), inicialmente discípulo de Watson, atacó sin piedad el conductismo dominante por entonces, estableciendo de modo solemne que la organización de la conducta no es impuesta desde el exterior sino que emana del interior del organismo.

Incluso un análisis superficial de este simposio de Pasadena de 1948 muestra aspectos interesantes. En primer lugar, el carácter interdisciplinar del programa de investigación que se ponía en marcha. Y en segundo lugar, la aparición de rasgos básicos en las ciencias cognitivas: la analogía entre computador y cerebro, el interés por los procesos de información y el anticonductismo que será tan típico de la psicología cognitiva.

Otro hito en la revolución cognitiva, también señalado por Gardner en el libro citado, fue el Simposio sobre Teoría de la Información que tuvo lugar en el famoso Instituto de Tecnología de Massachusetts (MIT) en 1956. Allen Newell y Herbert Simon, dos importantes informáticos, presentaron la primera demostración completa de un teorema realizada por un computador, en la línea de lo que se denominaría inteligencia artificial. A su vez, el lingüista Noam Chomsky presentó su enfoque transformacional de la gramática, que supone una actitud mentalista, frente al conductismo, insistiendo en que la organización del lenguaje "viene de dentro" y es universal. También, el psicólogo George Miller presentó un estudio sobre los límites humanos en la capacidad de procesar información, señalando que nuestra memoria inmediata se limita a aproximadamente siete items, aunque son posibles agrupaciones. Su célebre trabajo se tituló "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" (1956).

También aquí incluso un análisis superficial del simposio resulta revelador. Se muestra el carácter interdisciplinar del nuevo campo de investigación, así como rasgos básicos de las ciencias cognitivas: interés por los computadores, mentalismo, y concepción de la actividad mental como procesamiento de información.

Sin embargo, el comienzo institucional de la psicología cognitiva puede situarse en el año 1960, cuando los psicólogos George Miller y Jerome Bruner fundan el Harvard Center for Cognitive Studies en la Universidad de Harvard. En ese mismo año aparece el libro *Plans and the Structure of Behavior*, escrito por Miller, Galanter y Pribram, que constituye un auténtico manifiesto de la psicología cognitiva. Asimismo, en 1967 Ulric Neisser publica *Cognitive Psychology*, que es el primer libro de texto importante de la psicología cognitiva. Para Neisser, los procesos mentales humanos son más complejos que los programas de los computadores, por lo cual en vez de hablar de analogía del computador (para referirnos al parecido entre el procesamiento de información en los hombres y el procesamiento de información en un computador) prefiere hablar de analogía del programa, para indicar el hecho de que los seres humanos siguen un programa o conjunto de instrucciones al manipular la información.

2. Caracteres de la psicología cognitiva

La psicología cognitiva también puede denominarse psicología de procesamiento de información e incluso psicología computacional. Estos nombres atienden a la tesis básica de la psicología cognitiva de que los seres humanos (así como los animales superiores) al igual que los computadores son sistemas de procesamiento de información; dicho de modo más exacto, hombres (así como algunos animales) y computadores constituyen dos especies, con evidentes diferencias, del género que podemos llamar sistema de procesamiento de información. Una clara ilustración de esta asociación entre psicología cognitiva y procesamiento de la información es el contenido y el propio título del libro de R. Lachman, J. L. Lachman y E. C.

Butterfield: *Cognitive Psychology and Information Processing* (1979).

La psicología cognitiva se opone, tanto en su desarrollo histórico como en sus ideas centrales, al conductismo clásico, ya que postula la existencia de un sistema de estados internos en las criaturas inteligentes a fin de explicar su conducta. Es decir, entre los estímulos y respuestas, a los que se limita el conductismo clásico, se supone que existen procesos (que son precisamente los auténticos procesos mentales) que son causa de la conducta o bien funciones causales mediadoras entre estímulos y respuestas. Por ello, el mentalismo es esencial a la psicología cognitiva. Por otro lado, los estados internos postulados son portadores así como manipuladores de información, razón por la cual los procesos mentales o estados internos son procesos de cognición.

Margaret Boden, en su libro *Computer Models of Mind* (1988), señala tres modos característicos de teorizar compartidos por los psicólogos cognitivos o computacionales. En primer lugar, los psicólogos computacionales adoptan una aproximación funcionalista de la mente, en cuanto que los estados mentales se definen abstractamente en términos de su rol causal (con respecto a otros estados mentales y con respecto a la conducta observable). Se asume que cualquier fenómeno psicológico es generado por algún procedimiento efectivo, es decir, por algún conjunto de instrucciones, especificable de manera precisa, que define la sucesión de estados mentales dentro de la mente.

Respecto de este primer punto, me gustaría matizar que los procedimientos efectivos o algoritmos que siguen los procesos mentales no siempre son algoritmos en su sentido fuerte, esto es, procedimientos mecánicos sometidos a reglas fijas, sino que frecuentemente son algoritmos en sentido débil, esto es, ordenación racional de pasos caracterizada por la posibilidad de improvisar o completar algún paso, con lo que cabe un resultado variable dentro de un esquema genérico. Esto apunta, a mi entender, a la diferencia entre los procesos mentales humanos y los procesos mecánicos de los computadores clásicos, aunque ambos sean sistemas procesadores de

información, evitando tomar literalmente la comparación de la mente con el computador.

En segundo lugar, añade Boden, los psicólogos computacionales conciben la mente como un sistema representacional y consideran la psicología como el estudio de los diversos procesos computacionales según los cuales se construyen, organizan, interpretan y transforman las representaciones mentales. Como corolario, los psicólogos cognitivos usan una terminología "intencional" (o mejor dicho, me permito matizar, "representacional"), incluyendo gran parte del vocabulario propio de la psicología popular, es decir, piensan que los fenómenos mentales tienen un significado o contenido semántico y versan sobre algún objeto.

Y en tercer lugar, según Boden, los psicólogos computacionales o cognitivos se interesan en la neurociencia (cuando se interesan por ella) de una manera ampliamente computacional, preguntándose qué tipos de operaciones lógicas o relaciones funcionales podrían incorporarse en las redes neuronales. No importa el cerebro como un sistema físico en sí mismo, sino en cuanto incorpora las funciones mentales.

Estos tres modos de teorizar, que caracterizan la psicología cognitiva o computacional, constituyen una descripción general, pero dentro de la psicología cognitiva cabe distinguir dos grandes ramas, a saber, la psicología computacional clásica y la psicología computacional conexionista. Ambas ramas o tendencias básicas actuales se establecen según la relación de la psicología cognitiva con los modelos de computación empleados en la inteligencia artificial, ya que, como hemos visto en el capítulo 8, cabe distinguir dos modalidades básicas en los trabajos de inteligencia artificial, a saber, la inteligencia artificial clásica o GOFAI (esto es, la buena inteligencia artificial a la antigua usanza) y la inteligencia artificial conexionista.

Podemos caracterizar ambos tipos de psicología cognitiva de modo general, aunque luego haremos una caracterización más detallada. Para la psicología cognitiva clásica, los modelos de procesamiento de información responden a una computación consistente en manipulación de símbolos defini-

da en términos de reglas formales y realizada por un computador digital de propósito general con procesamiento serial, lo cual se denomina computador tipo von Neumann, en atención al célebre matemático e informático John von Neumann. En cambio, para la psicología cognitiva conexionista, los modelos de procesamiento de la información responden a una computación diseñada como manipulación de unidades con cierto grado de activación y que están masivamente interconectadas, de tal manera que se producen entre las unidades excitaciones e inhibiciones, y además tal computación diseñada sigue una pauta de procesamiento paralelo.

Dicho de otro modo, el modelo de computación que constituye el ideal de la psicología cognitiva clásica consiste en la aplicación serial o sucesiva de reglas formales almacenadas de manera explícita sobre representaciones simbólicas explícitas y localizables de manera definida. En cambio, el modelo de computación que constituye el ideal de la psicología cognitiva conexionista consiste en la aplicación en paralelo de reglas no predeterminadas de manera fija (sino flexibles) sobre unidades de activación conexionadas que representan al nivel de la red constituida por las unidades y no al nivel de cada unidad localizable.

Con todo, debemos establecer dos observaciones para una mejor comprensión. Por un lado, los modelos conexionistas de la psicología cognitiva conexionista suelen ser simulados de hecho en computadores tipo von Neumann (de procesamiento serial o sucesivo), aunque ya se han desarrollado arquitecturas típicamente paralelas, como, por ejemplo, en la obra de Hillis *The Connection Machine* (1985). Por otro lado, debe quedar claro que la psicología cognitiva en general concibe, tal como hemos señalado, los procesos mentales como procesos cognitivos (y de ahí su nombre), esto es, como procesos de recepción y manipulación de información. Por ello, la psicología cognitiva se esfuerza por enfocar sus investigaciones y presentar sus resultados en términos computacionales, usando los conceptos y modelos de la teoría de la computación. Sin embargo, éste es el ideal propuesto, ya que con frecuencia los resultados no llegan a presentar el aspecto "informático" acorde con aquel ideal.

3. Psicología cognitiva clásica

Tal como se acaba de indicar, para la psicología cognitiva clásica los tipos de procesos mentales son, o bien deben ser explicados como, tipos de computaciones formales del estilo de la inteligencia artificial clásica. Es decir, los procesos mentales son o se explican como manipulación de símbolos según programas de la clase de los empleados en los computadores de tipo von Neumann.

Ahora bien, esta asimilación de los procesos mentales con los procesos computacionales puede ser débil (metafórica) o bien fuerte (real). Según el primer punto de vista, la psicología cognitiva debe utilizar los conceptos y recursos computacionales para enfocar sus estudios y formular sus resultados, sin que se defienda que los procesos mentales son realmente computaciones iguales a las de un computador. A su vez, el segundo punto de vista sostiene que la mente realmente ejecuta computaciones iguales a las de un computador, es decir, la cognición en general se identifica con la computación mecánica.

Un autor representativo de este segundo punto de vista es el psicólogo Zenon Pylyshyn, tal como podemos advertir en su obra *Computation and Cognition* (1984). Pylyshyn señala que la posibilidad de imitar la vida mediante un artefacto ha intrigado a la gente a lo largo de la historia, pero que sólo en la segunda mitad del siglo XX se ha considerado seriamente la posibilidad de usar el especial tipo de artefacto llamado computador como un medio de comprender los fenómenos mentales, en cuanto procesos internos no (públicamente) observables. La extrema plasticidad de la conducta de los computadores es una razón por la que estos artefactos han sido considerados como capaces de exhibir inteligencia. Ha sido el fracaso en distinguir entre la computación como un tipo de proceso y la forma física particular que adopta en las máquinas de computación actuales lo que ha impedido a mucha gente tomar la computación como una explicación literal de los procesos mentales. Si comprendemos, concluye nuestro autor, la computación en un nivel completamente

general (como, de hecho, se comprende en teoría de la computación) podemos ver que la idea de que el procesamiento mental es computación es una hipótesis empírica seria y no una metáfora.

Ahora bien, el psicólogo que más esfuerzos ha dedicado a la justificación, explicación y defensa de la psicología cognitiva clásica es Jerry Fodor, del cual ya nos hemos ocupado en los capítulos 6 y 7. Para este autor, la psicología cognitiva o computacional clásica es la única psicología científica que puede llegar a realizarse. Sin embargo, Fodor ha mantenido dos posturas sucesivas respecto del carácter formal de los procesos mentales como manipulación de símbolos.

En efecto, en "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology" (1980), Jerry Fodor ha defendido una condición de formalidad que le lleva a un solipsismo metodológico, el cual, a su vez, pone en peligro la referencia de los procesos mentales a la realidad externa. Para nuestro autor, los procesos mentales son a la vez simbólicos y formales; los procesos mentales son simbólicos porque se definen sobre representaciones y son formales porque se aplican a las representaciones en virtud de la sintaxis de éstas. Pero esto último supone aceptar una condición de formalidad, consistente en que los estados mentales pueden ser de distinto tipo sólo si las representaciones que constituyen sus objetos son formalmente distintas. (Dicho de manera concreta, un estado mental es una percepción visual o es una imagen mental según el distinto carácter formal o computacional de sus representaciones).

Ahora bien, esta condición de formalidad implica el solipsismo metodológico. Hilary Putnam, en "The Meaning of 'Meaning'" (1975) había distinguido entre estados psicológicos en sentido estrecho (*narrow sense)* y estados psicológicos en sentido amplio (*wide sense*); los primeros son atribuidos a un sujeto sin presuponer ningún otro sujeto, mientras que los segundos son atribuidos a un sujeto presuponiendo la existencia de otro sujeto (por ejemplo, "x está celoso de y"). Putnam acuña la expresión "solipsismo metodológico" para indicar la exigencia de que sólo los estados psicológicos en sentido

estrecho son aceptados dentro de las teorías psicológicas. El propio Putnam considera que tal solipsismo metodológico constituye un programa restrictivo, pero Fodor, en el citado trabajo de 1980, acepta tal solipsismo metodológico como la estrategia de investigación propia de la psicología cognitiva. Para Jerry Fodor, los estados psicológicos estrechos son aquellos individualizados a la luz de la condición de formalidad, es decir, sin relación alguna con propiedades semánticas tales como verdad y referencia (al mundo externo).

Fodor, en su libro *Psychosemantics. The Problem of Meaning in the Philosophy of Mind* (1987), ha corregido, o bien completado, su oposición a la introducción de categorías semánticas en la psicología. El recurso empleado ha sido reivindicar la psicología popular o psicología del sentido común como parte de la psicología científica (tal como vimos en el capítulo 6). La psicología cognitiva recupera u obtiene la referencia al mundo externo a través de la psicología popular, que atribuye contenidos semánticos a los procesos mentales (al ser una psicología intencional). La explicación intencional es consubstancial con nuestro núcleo cognitivo, y los estados mentales no sólo tienen estructura sintáctica o formal sino también contenido.

4. Psicología cognitiva conexionista

Fodor y Pylyshyn unieron sus fuerzas para criticar la psicología conexionista, en su trabajo conjunto "Connectionism and cognitive architecture: A critical analysis" (1988). Sin embargo, no voy a comentar tales críticas, ni la respuesta de Paul Smolensky en "Connectionism, Constituency, and the Language of Thought" (1991). En lugar de ello, prefiero exponer las tesis positivas y la metodología especial propia de la psicología cognitiva conexionista. Mi punto de vista en la discusión entre psicólogos cognitivos clásicos y conexionistas es que ambos representan recursos y metodologías diferentes dentro de un mismo paradigma, a saber, el paradigma centrado en la idea de que los procesos mentales son estados internos (mentalismo) que manipulan información. Por otra parte,

entiendo que compete directamente a la inteligencia artificial evaluar ambas metodologías en cuanto técnicas y, a su vez, compete a la psicología cognitiva decidir la aplicación de una u otra según criterios de oportunidad.

Las redes neuronales, modelos conexionistas o, con un término más reciente, sistemas neuromórficos son modelos de computación construidos expresamente para reflejar los principios de organización y funcionamiento del cerebro humano. Este consta de muchas neuronas o células nerviosas, en torno a diez billones. Cada neurona recibe inputs o entradas de otras células, integra tales inputs y genera un output o salida, que envía a otras neuronas, o bien, en algunos casos, a órganos efectores como músculos o glándulas. Las uniones o sinapsis entre neuronas son de dos tipos generales, excitatorias e inhibitorias. Los neurocientíficos suelen medir la actividad de una neurona mediante su frecuencia de disparo. Las neuronas no son digitales, esto es, no tienen solamente estados de encendido o apagado en sus outputs, sino que tienen valores continuos en sus outputs, aunque a veces pueden ser analizadas como si fuesen digitales.

Pues bien, una red neuronal o modelo conexionista trata de reflejar esa estructura de las células nerviosas. Existen una serie de unidades simples de computación, con distintos valores de activación y con papeles de excitación o inhibición, que están masivamente conexionadas y operan en paralelo. Las unidades simples de la red se distribuyen en capas. Cuando hay sólo una capa de unidades tenemos un sistema autoasociativo; con dos capas una de ellas procesa el input del sistema y la otra proporciona el otput del sistema. Pero con frecuencia la red tiene varias capas de tal manera que, además de las capas externas de entrada y de salida del sistema, existe una o varias capas de unidades denominadas ocultas.

Las redes neuronales o modelos conexionistas se emplean tanto en inteligencia artificial como en psicología cognitiva, y además cabe señalar distintos grupos de investigación que los emplean, sin embargo, desde el punto de vista de la psicología cognitiva, el grupo más importante es el autode-

nominado grupo de investigación PDP (Parallel Distributed Processing), o grupo de procesamiento paralelo distribuido, con investigadores de la Universidad de San Diego (California), Universidad de Stanford y Universidad Carnegie-Mellon. Centraremos en este grupo nuestras consideraciones.

Los estudios de la aplicación de los modelos de procesamiento paralelo a los fenómenos psicológicos constituye el aspecto conexionista de este grupo PDP que comparte con otros grupos conexionistas. A su vez, el aspecto específico de este grupo es el empleo de representaciones distribuidas, en lugar de representaciones locales. Tal como señalan Geoffrey Hinton, James McClelland y David Rumelhart, miembros del grupo PDP, en su trabajo conjunto "Distributed Representations" (1986), dada una red de elementos simples de computación y dadas algunas entidades a representar, el esquema más directo de representación es usar un elemento de computación para cada entidad, y en tal caso tenemos una representación local. Pero, añaden estos autores, otro tipo de representación consiste en representar cada entidad por un patrón de activación distribuido entre muchos elementos de computación, con lo que cada elemento de computación está implicado en representar muchas entidades diferentes; y en este caso tenemos representaciones distribuidas. Conviene resaltar que existen grupos conexionistas, como el de Feldman y Ballard, de la Universidad de Rochester, que utilizan sistemas conexionistas locales.

Donald Norman, otro miembro del grupo PDP, en su trabajo "Reflections on Cognition and Parallel Distributed Processing" (1986), explica de modo lúcido las características, y novedades respecto de los modelos de computación clásica, de los modelos de procesamiento paralelo distribuido. En la computación clásica, el procesamiento se hace mediante componentes separados que se comunican pasando entre sí mensajes simbólicos, que se interpretan recurriendo a un sistema representacional que implica una correspondencia entre el símbolo y el referente y, por ende, un mecanismo interpretativo. Por tanto, el énfasis se pone en la representación simbólica. En cambio, en la computación conexionista PDP, el

procesamiento se hace mediante redes de procesamiento paralelo distribuido que se configuran a sí mismas para acomodarse a los datos que se reciben con discrepancia mínima. En consecuencia, los sistemas se ajustan continuamente, de modo que el aprendizaje es continuo, natural y fundamental. La información pasa entre las unidades de la red, no mediante mensajes, sino mediante valores de activación, es decir, empleando valores numéricos y no símbolos.

A su vez, Rumelhart, Hinton y McClelland, en su trabajo "A General Framework for Parallel Distributed Processing" (1986), enumeran los ocho aspectos básicos de un modelo PDP: 1) un conjunto de unidades de procesamiento, 2) un estado de activación del sistema, 3) una función de output para cada unidad, 4) un patrón de conectividad entre las unidades, 5) una regla de propagación para los patrones de actividades a través de la red de conectividades, 6) una regla de activación para combinar los inputs que afectan a una unidad con su estado actual para producir un nuevo nivel de activación para la unidad, 7) una regla de aprendizaje por la que los patrones de conectividad se modifican mediante la experiencia, y 8) un ambiente dentro del cual el sistema debe operar. Una explicación detallada de estos aspectos obligaría a la introducción del complejo aparato matemático de los modelos conexionistas.

Referencias bibliográficas

Aristóteles, *The Works of Aristotle* (J. A. Smith y W. D. Ross, comps.), Oxford, Clarendon Press, 1910-1952.
Armstrong, D., "The Nature of Mind", *Arts. The Proceedings of the Sydney University Arts Association*, 1966.
Armstrong, D., *A Materialist Theory of the Mind*, Londres, Routledge & Kegan Paul, 1968.
Baars, B., *A Cognitive Theory of Consciousness*, Nueva York, Cambrige University Press, 1988.
Boden, M., *Inteligencia Artificial y Hombre Natural* (trad. Julio C. Armero), Madrid, Tecnos, 1984.
Boden, M., *Computer Models of Mind*, Nueva York, Cambridge University Press, 1988.
Bunge, M. y Ardila, R., *Filosofía de la Psicología* (trad. Marco Aurelio Galmarini), Barcelona, Ariel, 1988.
Campbell, K. *Body and Mind* (ed. revis.), Notre Dame (Indiana), University of Notre Dame Press, 1984.
Carnap, R., "Psicología en lenguaje fisicalista", A. J. Ayer (comp.), *El positivismo lógico* (trad. L. Aldama y otros), México, Fondo de Cultura Económica, 1965.
Cordemoy, G. D. E., *Discurso Físico de la Palabra* (trad. Pedro J. Chamizo), Málaga, Universidad de Málaga, 1989.
Chomsky, N., "A Review of B. F. Skinner's *Verbal Behavior*", *Language*, 35 (1959), pp. 26-58.
Churchland, P. M. "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy*, 78, nº 2 (febrero 1981), pp. 67-90.
Churchland, P. M., *Matter and Consciousness* (ed. revis.), Cambridge (Mass.), MIT Press, 1988. [*Materia y conciencia* (trad. Margarita Mizraji), Barcelona, Gedisa, 1992.]
Churchland, P. M. y Churchland, P. S., "Could a Machine Think?", *Scientific American*, vol. 262, nº 1 (enero 1990), pp. 32-37.

Descartes, *Oeuvres et Lettres* (André Bridoux, comp.), París, Gallimard, 1953.
Dreyfus, H., *What Computers Still Can't Do*, Cambridge (Mass.), MIT Press, 1992.
Dreyfus, H. y Dreyfus, S., *Mind over Machine*, Oxford, Basil Blackwell, 1986.
Eccles, J., *The Neurophysiological Basis of Mind: The Principles of Neurophysiology*, Oxford, Clarendon Press, 1953.
Eccles, J., "Cerebral Activity and Consciousness", F. J. Ayala y T. Dobzhansky (comps.), *Studies in the Philosophy of Biology*, Berkeley, University of California Press, 1974.
Eccles, J., *La psique humana* (trad. Carmen García Trevijano), Madrid, Tecnos, 1986.
Eccles, J., *Evolution of the Brain: Creation of the Self*, Londres, Routledge, 1989.
Eccles, J. y Zeier, H., *El cerebro y la mente* (trad. Diorki), Barcelona, Herder, 1985.
Fodor, J., *La explicación psicológica* (trad. José E. García Albea), Madrid, Cátedra, 1980.
Fodor, J., "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", *The Behavioral and Brain Sciences*, 3 (1980), pp. 63-73.
Fodor, J., *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge (Mass.), MIT Press, 1987.
Fodor, J., y Pylyshyn, Z., "Connectionism and cognitive architecture: A critical analysis", *Cognition*, 28 (1988), pp. 3-71.
Gardner, H., *La nueva ciencia de la mente* (trad. Leandro Wolfson), Barcelona, Paidós, 1988.
Gil de Pareja, J. L., *La Filosofía de la Psicología de Ludwig Wittgenstein*, Barcelona, PPU, 1992.
Haugeland, J., *Artificial Intelligence. The Very Idea*, Cambridge (Mass.), MIT Press, 1985.
Hillis, W. D. *The Connection Machine*, Cambridge (Mass.), MIT Press, 1985.
Hinton, G., McClelland y Rumelhart, D., "Distributed Representations", D. Rumelhart, J. McClelland and the PDP Research Group, *Parallel Distributed Processing*, vol. 1, Cambridge (Mass.), MIT Press, 1986.
Jastrow, R., *El telar mágico* (trad. Domingo Santos), Barcelona, Salvat, 1988.
Kornhuber, H. H., "Cerebral Cortex, Cerebellum and Basal Ganglia: An Introduction to their Motor Functions", F. O. Schmitt y F. G. Worden (comps.), *The Neurosciences*, Third Study Program, Cambridge (Mass.), MIT Press, 1974.

Lachman, R., Lachman, J. L., y Butterfield, E. C., *Cognitive Psychology and Information Processing*, Hillsdale, Lawrence Erlbaum, 1979.

Lashley, K., *Brain Mechanisms and Intelligence*, Chicago, University of Chicago Press, 1929.

Lewis, D., "An Argument for the Identity Theory", *Journal of Philosophy*, 63 (1966), pp. 17-25.

Lewis, D., "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy*, 50 (1972), pp. 249-258.

Libet, B., "Electrical Stimulation of Cortex in Human Subjects, and Conscious Sensory Aspects", A. Iggo (comp.), *Handbook of of Sensory Physiology*, vol. II, Berlín, Springer, 1973.

Lukaszewicz, W., *Non-Monotonic Reasoning*, Londres, Ellis Horwood, 1990.

Machine Learning (Special Issue on Reinforcement Learning), vol. 8, nº 3/4 (1992).

Martínez-Freire, Pascual F., "Antimentalismo y mentalismo en las *Investigaciones Filosóficas* de Wittgenstein", *Arbor*, nº 532 (abril de 1990), pp. 79-90.

Martínez-Freire, P. F., "Procesos mentales y cognitivismo", *Revista de Filosofía*, vol. V, nº 7 (1992), pp. 143-159.

Martínez-Freire, P. F., "Delimitación de las ciencias cognitivas", *Anales del Seminario de Metafísica*, nº extra (1992), pp. 443-451.

Miller, G., "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information", *Psychological Review* 63 (1956), pp. 81-97.

Miller, G., Galanter, E., y Pribram, K., *Plans and the Structure of Behavior*, Nueva York, Holt, Rinehart and Winston, 1960.

Neisser, U., *Cognitive Psychology*, Nueva York, Meredith, 1967.

Norman, D., "Reflections on Cognition and Parallel Distributed Processing", J. McClelland, D. Rumelhart and the PDP Research Group, *Parallel Distributed Processing*, vol. 2, Cambridge (Mass.), MIT Press, 1986.

Penfield, W., *El misterio de la mente* (trad. Santiago Páez Fuentes), Madrid, Pirámide, 1977.

Penrose, R., *La nueva mente del emperador* (trad. Javier García Sanz), Madrid, Mondadori, 1991.

Platón, *El Banquete, Fedón y Fedro* (trad. Luis Gil), Labor, Barcelona, 1981 (4ª ed.).

Popper, K., *Conocimiento objetivo* (trad. Carlos Solís Santos), Madrid, Tecnos, 1974.

Popper, K. y Eccles, J., *El yo y su cerebro* (trad. C. Solís Santos), Barcelona, Labor, 1980.

Putnam, H., "Psychological Predicates", W. H. Capitan y D. D. Merrill

(comps.), *Art, Mind and Religion*, Pittsburgh, University of Pittsburgh Press, 1967. (Reimpreso posteriormente con el título de "The Nature of Mental States").
Putnam, H., "Reductionism and the Nature of Psychology", *Cognition*, 2 (1973), pp. 131-146.
Putnam, H., "The Meaning of 'Meaning'", K. Gunderson (comp.), *Language, Mind, and Knowledge*, Minnesota Studies in the Philosophy of Science, vol. VII, Minneapolis, University of Minnesota Press, 1975.
Putnam, H., *Reason, Truth and History*, Nueva York, Cambridge University Press, 1981.
Putnam, H., *Representación y realidad* (trad. Gabriela Ventureira), Barcelona, Gedisa, 1990.
Pylyshyn, Z., *Computation and Cognition*, Cambridge (Mass.), MIT Press, 1984.
Rapaport, W., "Syntactic Semantics: Foundations of Computational Natural-Language Understanding", J. Fetzer (comp.), *Aspects of Artificial Intelligence*, Dordrecht, Kluwer, 1988.
Raphael, B., *The Thinking Computer*, Freeman, San Francisco, 1976.
Rumelhart, D., Hinton, G. y McClelland, J., "A General Framework for Parallel Distributed Processing", D. Rumelhart, J. McClelland and the PDP Research Group, *Parallel Distributed Processing*, vol. 1, Cambridge (Mass.), MIT Press, 1986.
Russell, B., *The Analysis of Mind*, Nueva York, Humanities Press, 1978.
Ryle, G., *El Concepto de lo Mental* (trad. Eduardo Rabossi), Buenos Aires, Paidós, 1967.
Searle, J., "Minds, Brains, and Programs", *The Behavioral and Brain Sciences*, 3 (1980), pp. 417-424.
Searle, J., *Mentes, cerebros y ciencia* (trad. Luis Valdés), Madrid, Cátedra, 1985.
Searle, J., "Minds and Brains Without Programs", C. Blakemore y S. Greenfield (comps.), *Mindwaves*, Oxford, Basil Blackwell, 1987.
Searle, J., *The Rediscovery of the Mind*, Cambridge (Mass.), MIT Press, 1992.
Skinner, B. *The Behavior of Organisms*, Nueva York, Appleton-Century-Crofts, 1938.
Skinner, B., *Walden Dos*, Barcelona, Ediciones Martínez Roca, 1984.
Skinner, B., *Science and Human Behavior*, Nueva York, Macmillan, 1953.
Skinner, B., *Verbal Behavior*, Nueva York, Appleton-Century-Crofts, 1957.
Skinner, B., "Behaviorism at fifty", T. W. Wann (comp.), *Behaviorism*

and Phenomenology, Chicago, University of Chicago Press, 1964.
Skinner, B., *Beyond Freedom and Dignity*, Nueva York, Knopf, 1971.
Skinner, B., *About Behaviorism*, Nueva York, Random House, 1974.
Smolensky, P., "Connectionist AI, Symbolic AI, and the Brain", *Artificial Intelligence Review*, 1(1987), pp. 95-109.
Smolensky, P., "Connectionism, Constituency, and the Language of Thought", B. Loewer y G. Rey (comps.), *Meaning in Mind. Fodor and his Critics*, Oxford, Basil Blackwell, 1991.
Sperry, R., "Mind-Brain Interaction: Mentalism, yes; Dualism, no", *Neuroscience*, vol. 5 (1980), pp. 195-206.
Sperry, R. W., Gazzaniga, M. S. y Bogen, J. E., "Interhemispheric Relationships: the Neocortical Commisures: Syndromes of Hemisphere Disconnection", P. J. Vinken y G. W. Bruyn (comps.) *Handbook of Clinical Neurology*, vol. 4, Amsterdam, North Holland, 1969.
Stich, S., "Autonomous Psychology and the Belief-Desire Thesis", *Monist*, 61 (1978), pp. 573-591.
Stich, S., *From Folk Psychology to Cognitive Science. The Case against Belief*, Cambridge (Mass.), MIT Press, 1983.
Szentágothai, J., "The Basic Neuronal Circuit of the Neocortex", H. Petsche y M. A. B. Brazier (comps.), *Synchronization of EEG Activity in Epilepsies*, Berlín, Springer, 1972.
Tolman, E., *Purposive Behavior in Animals and Men*, Nueva York, Appleton-Century-Crofts, 1932.
Turing, A., "On Computable Numbers, with an Application to the *Enstscheidungsproblem*", *Proceedings of the London Mathematical Society*, ser. 2, 42 (1936-37), pp. 230-265.
Turing, A., "Computing Machinery and Intelligence", *Mind*, 59 (1950), pp. 434-460.
Watson, J., "Psychology as a Behaviorist views it", *Psychological Review*, 20 (1913), pp. 158-177.
Watson, J., *Behavior: An Introduction to Comparative Psychology*, Holt, Nueva York, Rinehart and Winston, 1914.
Watson, J., *Psychology from the Standpoint of a Behaviorist*, Philadelphia, Lippincott, 1919.
Watson, J., *Behaviorism*, Nueva York, People's Institute, 1925.
Weizenbaum, J. "ELIZA. A Computer Program for the Study of Natural Language Communication between Man and Machine", *Communications of the Association for Computing Machinery*, 9 (1965), pp. 36-45.
Weizenbaum, J., *Computer Power and Human Reason*, San Francisco, Freeman, 1976.
Whitehead, A. N. y Russell, B. *Principia Mathematica* (hasta *56)

(trad. J. Manuel Domínguez Rodríguez), Madrid, Paraninfo, 1981.
Wittgenstein, L., *Tractatus Logico-Philosophicus* (trad. Jacobo Muñoz e Isidoro Reguera), Madrid, Alianza, 1987.
Wittgenstein, L., *Zettel* (trad. G. E. M. Anscombe), Oxford, Basil Blackwell, 1981 (2ª ed.).
Wittgenstein, L., *Investigaciones Filosóficas* (trad. Alfonso García Suárez y Ulises Moulines), Barcelona, Crítica, 1988.
Wundt, W., *Eléments de psychologie physiologique* (trad. E. Rouvier), París, Alcan, 1886.

Indice de autores

Ardila, Rubén 61
Aristóteles 18, 19, 28
Armstrong, David 14, 78-81, 82, 85
Arnauld, Antoine 29
Austin, John 42

Baars, Bernard 87, 88
Ballard, Dana 153
Boden, Margaret 146, 147
Bogen, Joseph 136
Bruner, Jerome 145
Bunge, Mario 61
Butterfield, Earl 146

Campbell, Keith 84, 96
Carnap, Rudolf 21
Cordemoy, Gerauld de 26

Chomsky, Noam 61, 144
Churchland, Patricia 115, 116-7
Churchland, Paul 69, 70, 71, 72-3, 115, 116-7

Descartes, René 14, 19-20, 25-40, 80, 124, 125
Dreyfus, Hubert 105-110, 119
Dreyfus, Stuart 105, 107, 108

Eccles, John 15, 125, 126, 132-42

Feldman, Jerome 153
Feyerabend, Paul 79
Fodor, Jerry 14, 66-9, 78, 94-6, 150, 151

Galanter, Eugene 62, 145
Galeno 28
Galvani, Luigi 28
Gardner, Howard 143, 144
Gazzaniga, Michael 136
Gil de Pareja, José Luis 158

Haugeland, John 108, 116
Heidegger, Martin 107
Hillis, Daniel 148
Hinton, Geoffrey 153, 154
Hipócrates 28
Hull, Clark 62
Hume, David 37

James, William 21
Jastrow, Robert 119

Kornhuber, Hans Helmut 138, 139, 140

Lachman, Janet 145
Lachman, Roy 145
Lakatos, Imre 109
Lashley, Karl 61, 144

Lederberg, Joshua 104
Leibniz, Gottfried Wilhelm 20, 32
Lewis, David 14, 78, 82-3, 85, 86
Libet, Benjamin 138, 140
Locke, John 97
Lukaszewicz, Witold 121

Martínez-Freire, Pascual F. 159
McCarthy, John 103
McClelland, James 153, 154
McCulloch, Warren 144
Merleau-Ponty, Maurice 107
Miller, George 62, 144, 145
Minsky, Marvin 103, 105, 109

Neisser, Ulric 145
Newell, Allen 103, 109, 144
Norman, Donald 153

Oettinger, Anthony 104
Ortega y Gasset, José 88

Penfield, Wilder 125, 127, 128
Penrose, Roger 140
Platón 18, 19
Popper, Karl 132, 139, 140
Pribram, Karl 62, 145
Putnam, Hilary 14, 74, 78, 89-94, 97, 98, 127, 150-1
Pylyshyn, Zenon 149, 151

Rapaport, William 115
Raphael, Bertram 105
Roberts, Larry 104
Rumelhart, David 153, 154
Russell, Bertrand 22, 103
Ryle, Gilbert 13, 14, 22, 31-40, 42, 53, 80

San Agustín 124, 125
Samuel, Arthur 103
Schlick, Moritz 21
Searle, John 15, 112-18, 126, 128-32
Shaw, Cliff 103
Sherrington, Charles 132
Simon, Herbert 103, 109, 144
Skinner, Burrhus 57, 60, 61
Smolensky, Paul 110, 151
Spence, Kenneth 62
Sperry, Roger 132, 135, 136, 139, 140
Spinoza, Benito 20, 25
Stich, Stephen 69, 70
Strawson, Peter 42
Szentágothai, Janos 134

Thorndike, Edward 54
Tolman, Edward 62
Turing, Alan 91, 99-101, 102, 112, 114, 122

Von Neumann, John 144, 148, 149

Watson, John 38, 46, 54, 55-6, 57, 61, 144
Weizenbaum, Joseph 110-2, 113, 120
Whitehead, Alfred 103
Wittgenstein, Ludwig 13, 14, 22, 32, 41-51, 53, 107, 118
Wundt, Wilhelm 21

Zeier, Hans 134
Zuse, Konrad 102